Vol. 69, No. 6 November 1962

Psychological Review

RICHARD L. SOLOMON, Editor
University of Pennsylvania

CONTENTS

ARTICLES

Observations on Active Touch.....JAMES J. GIBSON 477

A System for Shape Recognition.....J. A. DEUTSCH 492

Patterns of Adrenergic-Cholinergic Imbalance in the Functional Psychoses.....LEONARD S. RUBIN 501

Success and Speed of Problem Solving by Individuals and Groups
FRANK RESTLE AND JAMES H. DAVIS 520

THEORETICAL NOTES

The Molecular Basis of Memory and Learning
MICHAEL H. BRIGGS AND G. BARRIE KITTO 537

The Untenability of Luce's Principle.....WILLIAM W. ROZEBOOM 542

Comments on Rozeboom's Criticism of "On the Possible Psychophysical Laws".....R. DUNCAN LUCE 548

Comment.....WILLIAM W. ROZEBOOM 552

This is the last issue of Volume 69.
Title page and index for the volume
appear herein.

PUBLISHED BIMONTHLY BY THE
AMERICAN PSYCHOLOGICAL ASSOCIATION, INC.

Editorial Consultants

GORDON W. ALLPORT
JUSTIN M. ARONFREED
ABRAHAM H. BLACK
DONALD M. BROVERMAN
RAYMOND B. CATTELL
JOHN GAITO
EUGENE GALANTER
DAVID H. GRANT
G. ROBERT GRICE
R. J. HERRNSTEIN
MICHAEL KAHN
LEON J. KAMIN
GARDNER LINDZEY
R. DUNCAN LUCE
ROBERT B. MACLEOD
GEORGE MANDLER
SARNOFF A. MEDNICK
O. HOBART MOWRER
CONRAD G. MUELLER
ALBERT PEPITONE
CHARLES C. PERKINS, JR.
CARL PFAFFMANN
MARK R. ROSENZWEIG
BURTON S. ROSNER
WILLIAM W. ROZEBOOM
HAROLD SCHLOSBERG
ROGER N. SHEPARD
JOSEPH SHOBEN
MONCRIEFF H. SMITH
KARL U. SMITH
SAUL STERNBERG
PATRICK SUPPES
WARREN S. TORGERSON
WILLIAM S. VERPLANCK
DAVID R. WILLIAMS

(A large group of psychologists has kindly agreed to assist in the editing of manuscripts submitted to the *Psychological Review*. In each issue the Editor will list in the *Review* those editorial consultants who have helped him process manuscripts for that issue, so that their efforts may be gratefully acknowledged.)

The *Psychological Review* is devoted to articles of theoretical significance to any area of scientific endeavor in psychology. Except for occasional articles solicited by the Editor, manuscripts exceeding 7,500 words will not be accepted. Ordinarily manuscripts which consist primarily of original reports of research should be submitted to other journals. However, reports of original research may be included in a major theoretical article which attempts to integrate several related original studies.

Because of the large number of manuscripts submitted, there is an inevitable publication lag of several months. Authors may avoid this delay if they are prepared to pay the costs of publishing their own articles; the appearance of articles by other contributors is not thereby delayed.

Manuscripts must be accompanied by an abstract of 100 to 120 words typed on a separate sheet of paper. The abstract should conform to the style of *Psychological Abstracts*. Detailed instructions for preparation of the abstract may be obtained from the Editor or from the APA Central Office.

All manuscripts must be submitted in duplicate.

Original figures are prepared for publication; duplicate figures may be photographic or pencil-drawn copies. Authors should retain a copy of the manuscript, in case there is loss in the mail. Manuscripts should be addressed to the Editor, Richard L. Solomon, Psychology Department, 5 College Hall, University of Pennsylvania, Philadelphia 4, Pennsylvania.

HELEN ORR, Managing Editor
ELIZABETH S. REED, Advertising Manager
MARY SUSAN GILBERT, Editorial Assistant

PUBLISHED BIMONTHLY BY THE
AMERICAN PSYCHOLOGICAL ASSOCIATION, INC.
Prince and Lemon Streets, Lancaster, Pa.
1333 Sixteenth St. N.W., Washington 6, D. C.

$10.00 volume $2.00 issue

Second-class postage paid at Lancaster, Pa.

Send all communications, including address changes, to 1333 Sixteenth St. N.W., Washington 6, D. C. Address changes must arrive by the tenth of the month to take effect the following month. Undelivered copies resulting from address changes will not be replaced; subscribers should notify the post office that they will guarantee second-class forwarding postage. Other claims for undelivered copies must be made within four months of publication.

© 1962 by the American Psychological Association, Inc.

VOL. 69, NO. 6 NOVEMBER 1962

PSYCHOLOGICAL REVIEW

OBSERVATIONS ON ACTIVE TOUCH[1]

JAMES J. GIBSON

Cornell University

Active touch refers to what is ordinarily called *touching*. This ought to be distingushed from passive touch, or *being touched*. In one case the impression on the skin is brought about by the perceiver himself and in the other case by some outside agency. The difference is very important for the individual but it has not been emphasized in sensory psychology, and particularly not in the experimental literature.

Active touch is an exploratory rather than a merely receptive sense. When a person touches something with his fingers he produces the stimulation as it were. More exactly, variations in skin stimulation are caused by variation in his motor activity. What happens at his fingers depends on the movements that he makes—and also, of course, on the object that he touches. Such movements are not the ordinary kind usually thought of as responses. They do not modify the environment but only the stimuli coming from the environment. Presumably they enhance some features of the potential stimulation and reduce others. They are exploratory instead of performatory. In this respect, these touching movements of the fingers are like the movements of the eyes. In fact, active touch can be termed *tactile scanning*, by analogy with ocular scanning.

By means of active touch a great many properties of the adjacent environment can be perceived in the absence of vision. The blind must depend on it for most of their information about the world (Revesz, 1950). Nevertheless, despite its importance, the "sense of touch" (Boring, 1942, Ch. 13; Geldard, 1953, Ch. 9–12) has been studied by sensory physiologists only as a passive or receptive channel. It is treated as a part of cutaneous sensitivity. The sensitivity of the skin, which includes temperature and pain as well as touch, can be most easily investigated by the applying of stimuli to the cutaneous surface. At a perceptual level, Geldard (1957) has applied multiple stimuli to the skin and has thereby demonstrated some of the various "messages" that the skin will transmit but he is nevertheless concerned with a mosaic of receptors, not an exploratory organ. Katz (1925) and Revesz (1950) have argued, however, that the *hand* is a kind of sense organ as distinguished from the skin of the hand. Katz began the description of tactual experience as it occurs in life, and performed experiments on

[1] This research was supported in part by the Office of Naval Research under Contract NONR 401 (14) with Cornell University.

some of the remarkable discriminations that are possible. Revesz, observing the performances of the blind, proposed an unrecognized mode of experience called *haptics* which goes beyond the classical modalities of touch and kinesthesis. His interest is in philosophy and esthetics. These two seem to be almost the only investigators who have given much thought to what is here called active touch. Their work has not been followed up by experimenters and the term itself has never been widely used, perhaps because it does not fit with the established ideas of what constitutes a sense.

The relation of touch to kinesthesis. It seems to have been assumed by psychologists that active touch is simply a blend of two modes of sensation, kinesthesis and touch proper. That is, touching merely combines the data from the feeling of movement and the feeling of contact, the two sensations being fused into one experience. This assumption, however, can be criticized. First, it fails to take account of the purposive character of touching and, second, it fails to emphasize the multiplicity of so-called kinesthesis.

The first objection is that the act of touching or feeling is a search for stimulation or, more exactly, an effort to obtain the kind of stimulation which yields a perception of what is being touched. When one explores anything with his hand, the movements of the fingers are purposive. An organ of the body is being adjusted for the registering of information. The limbs and extremities are, of course, motor organs as well as sensing organs, whereas the eyes are only sense organs, but the function of motor performance can be subordinated to the function of exploratory adjustment in the case of the limbs.

Behavior theorists have recently been emphasizing the importance of what is variously called "feedback" from responses, or self-produced stimulation, or proprioception in general. They recognize that such inputs are necessary for the purposive control or steering of behavior. What they have not recognized is that some of the stimulation produced by responses yields objective information as well as subjective. The stimulation produced by sense-organ adjustment is of this sort. In active touch, as will later be suggested, the flow of stimulation contains two components, one *exterospecific* and one *propriospecific*. But these components are not the traditional sensations of touch and kinesthesis respectively. The hypothesis will be advanced that the purpose of the exploratory movements of the hand is to isolate and enhance the component of stimulation which specifies the shape and other characteristics of the object being touched.

A second objection is that the term kinesthesis itself means different things and should not continue to imply the existence of a unitary sense. Historically the word refers to movement of the body, and originally meant the "muscle sense." But it is known that there is sensitivity to the position of the body and of all its members relative to one another. This is an articular sense, not a muscle sense, and the joints yield information about joint position as well as joint rotation (Goldscheider, 1898). Probably there is no "sense" of muscle contraction as such, the joint initiating the spatial information and the muscle spindles having only reflex coordinating functions (Rose & Mountcastle, 1959). There is sensitivity for the upright posture of the head, the linear and angular accelerations of the head, and a general "sense" of the equilibrium of the body. There is sensitivity of some kind for the force exerted by an

individual, whether or not this is accompanied by hypothetical "feelings of innervation" (Boring, 1942, Ch. 14). There is also, of course, *visual* sensitivity to the changes of position of one's body in space. And there is proprioception in general, however this may be defined, as well as somaesthesia.

The single term "kinesthesis" cannot carry the burden of all the meanings that have been added to its original meaning. It fails to do justice to different inputs, to different combinations of these inputs, and to different functions of these combinations.

Types and subtypes of anatomical receptors involved in active touch. Passive touch involves only the excitation of receptors in the skin and its underlying tissue, although the patterns may be elaborate. But active touch involves the concomitant excitation of receptors in the joints and tendons along with new and changing patterns in the skin. Moreover, when the hand is feeling an object, the movement or angle of each joint from the first phalanx of each finger up to the shoulder and the backbone makes its contribution. And these inputs occur relative to a continuous input from the vestibular organs, along with the cutaneous input from contact with the ground. Presumably the feeling of an object by the hand involves the feeling of the position of the fingers, hand, arm, body, and even the head relative to gravity, all being integrated in some hierarchy of positional information.

The total flux of stimulation is enormously complex, but lawful modes of combination occur. Presumably the modes of combination of these inputs specify the difference between touching and being touched. Also certain special combinations probably carry information about the *object* being touched. In short, the so-called sense of touch as it is most often employed by using the hands involves the play of inputs coming from the whole skeleto-muscular system. Whether it is one sense or many is therefore a matter of dispute. It depends on what one wishes to call a sense. There is no single organ or structure analogous to an ear or an eye. There is no unique quality of sensation analogous to warm, cold, pain, sour, sweet, or red. Active touch does not fulfill the supposed criteria for a single sense modality. Nevertheless it provides a quite definite channel of information about the external environment. It is a type of perception that is isolable from vision, audition, taste, and smell, and it needs to be studied in its own right.

Mechanics of cutaneous stimulation, proximal and distal. The literal or proximal stimulus for touch, the immediate cause, is some sort of deformation of the skin, not necessarily a simple depression of the surface as by a tactile "stimulator," but almost any deformation of the elastic tissue. Pulling on a thread glued to the skin is just as effective a stimulus as pushing it in. The stimulus has to be a change in time, that is, a motion of the skin. Since the work of Nafe and Wagoner (1941) it has been suspected that whenever the skin becomes absolutely motionless even in some abnormal conformation, the receptors cease firing and the stimulus becomes ineffective. Outside of the laboratory this situation would be rare. A precisely constant pressure of an object relative to the skin would not occur, if only by reason of tremor or body sway.

The cause of elastic change of the skin's surface can be almost any mechanical event. Some work must be done, even if slight, but the actual me-

chanical happenings are so various and complex that the textbooks of mechanics do not begin to describe them. A common event is impact with a solid substance, but a puff of air or a drop of water will do as well. In hairy regions of the skin, even the slightest bending of a hair gives a stimulus. Some of the types of mechanical events that stimulate the skin, in terms of everyday physics, may be exemplified as follows.

1. Brief events: pressure, push, slap, pat, tap, prick. Note that these vary in area as well as in duration.

2. Prolonged events without displacement: vibration, stretching, kneading, pinching.

3. Prolonged events with displacement: scratching, scraping, rubbing, sliding, brushing, rolling. Note that these moving deformations vary in the degree of friction between the two surfaces, and also in the degree of lateral stretching of the skin as well as the degree of vertical depression.

These are relatively large scale events as compared to the micro-events that excite the actual end-organs imbedded in the tissue. But the latter are also mechanically complex. The ways in which mechano-receptors are made to fire nervous impulses need not be considered here. The facts, so far as they are known, are to be found in the physiological literature.

The external causes of these mechanical events, the sources of potential cutaneous stimulation, are the objects and surfaces of the environment. They are at the distal end of a sequence that leads to the excitation of nerve cells. An objective surface of some kind and the cutaneous surface must "touch," either by movement of an object with reference to the stationary skin or by movement of the skin with reference to a stationary object.

Contact with the substratum. The "sense of touch" is not ordinarily thought to include the feeling of cutaneous contact with the earth. Nevertheless, the upward pressure of the surface of support on some part of the body provides, for every terrestrial animal, a constant background of stimulation. It is covariant with the continuous input of the statocyst receptors of the inner ear already mentioned. Together they provide what the ordinary person calls the "sense of support." The axis of gravity and the plane of the ground provide the basic frame of reference for tactual space perception. Active touch, as will appear, yields clear perceptions of environmental space in the absence of vision.

The touching of the substratum is neither active nor passive touch as defined above, i.e., neither the touching of a stationary object nor being touched by a moving object. It is instead a means of registering the stable environment with reference to which *both* the movements of one's body and the motions of objects occur.

OBSERVATIONS OF THE ACT OF TOUCHING [2]

In the simple experiments to be described, vision was always excluded. This was usually accomplished by having the seated observer put both hands under a cloth curtain at the front of a table, not by blindfolding him, or working in darkness. His manual activity could thus be observed.

When an object of any sort is placed in the hand of the observer or when

[2] The collaborators in this work were Susan Stanley, Lewis Greenberg, and James Caviness. The report has profited from a critical reading by Jesse Smith.

his hand is placed on it, the following facts can be noted. He tends to bring the other hand to the object if allowed to do so. He tends strongly to curve his fingers around the object if permitted. He tends to move his fingers over the object in complex ways. Some of these are: tracing movements with one fingertip or several fingertips, opposition of the thumb and other fingers, and rubbing, grasping, or pressing movements of the fingers. Finally, he tends to name the object if he can do so, or to compare it with a familiar object if he cannot. Nonsense objects are almost necessary for such observations since the observer does not persist after recognizing something.

The observer seems to be trying to obtain mechanical events at the skin at various places in various combinations. The movements by which he obtains them do not ever seem to be twice the same, but they are not aimless. If the hand is conceived as a sense organ, he seems to be adjusting it. He appears to be searching for stimulus information.

So much for the manual activity. Reports of experience during active touch can also be obtained. They tend to be radically different from reports of experience during passive touch. The following facts, noted by the writer, are confirmed by other observers in this experiment, or are consistent with earlier observations.

Unity of the phenomenal object. When feeling a single object with two fingers, only one object is perceived although two separated cutaneous pressures occur. The separate "local signs" are not noticed. This holds true not only for finger-thumb opposition but also for other combinations of fingers. The unitary perception occurs when all five fingers are applied to an object, and even when two separate hands are applied to the object. In fact, 10 different digital pressures all at the same time yield a wholly unified experience. There are some limitations on this unity, as can be observed with Aristotle's illusion when a pencil is held between crossed fingers, but even then it is very hard to overcome, by the writer's introspection. This fact needs further investigation.

Stability of the phenomenal object. When sliding the skin over a corner or protuberance of an object, the displacement of the cutaneous pressure, the "tactile motion," cannot usually be noticed. The object seems to remain stationary even though the impression moves relative to the skin. Not only the object but also the whole space of the table, chair, floor, and room is felt as perfectly stable. This phenomenon has been studied in an experiment reported in the next section.

Rigidity or plasticity of the phenomenal object. When pressing a finger on a rigid surface or squeezing an object with the hand, it is difficult to notice the increase of intensity of cutaneous sensation; instead the observer is primarily aware of the substance and its resistance. Likewise, when pressing or squeezing a nonrigid object (a lump of modeling clay, a rubber ball, a handful of cloth) the observer is aware of the yieldingness, elasticity, or softness of the substance, not of the (very different) intensity of the back pressure on the skin. This fact deserves further study. Presumably the degree of force exerted is being registered by the joint-and-tendon receptors and the resulting degree of skin-pressure is being registered by the cutaneous and deep-tissue receptors, but the intensities are not experienced as such. Perhaps it is the *relation* between these different intensities that is registered. The proportion of

one to the other, and the relative temporal buildup of these two, are different for rigid and for nonrigid objects. A possible hypothesis is that proportion and buildup are registered by the receptive system when it is operating in an exploratory fashion. They constitute stimulus-information about the object, whereas the intensities as such do not.

Shape of the phenomenal object. When the corners, edges, or other protuberances of a strange object are being felt, one can distinguish the pattern which these make to one another but one cannot distinguish the pattern which the various cutaneous pressures make to one another. One perceives the object-form but not the skin-form. The latter is, in fact, continually changing as the fingers move in various ways. It is almost completely unreportable, whereas the pattern of physical corners and edges seems to emerge in experience. This fact is extremely interesting and two experiments arising from it will later be reported.

Discussion. For all four of the perceptions described above, it is possible to perform a sort of control experiment substituting passive touch for active touch. The experiences aroused are quite different. For the first, two different pressures applied to two different loci on the skin of the hand yield two sensory impressions, not one. For the second, translocation of the stimulus over the skin yields a perception of something moving. For the third, increase in the intensity of skin-depression is reportable as such. For the fourth, a pattern of simultaneous pressures is detectable when it is pushed into the passive skin. In all cases the sensory impressions can be aroused by an experimenter, but when the observer himself brings them about they seem to disappear.

Different Ways of Perceiving Motion of an Object by Touch

The objects of tactile perception considered so far have been essentially stationary, either resting on a table top or in one or both hands. How can the motion of an object be perceived, and how can this be distinguished from the motion of the hand relative to the object?

Translatory motion over the skin. The simplest case is that when the observer's hand lies passively supported behind the curtain and an object of some sort is made to trace a path of pressure across the skin. The path, direction, and speed of motion could always be described relative to the skin. They could equally well be described relative to the table top and to the space of the room.

For these observations various instruments were used, a sharpened point, a blade, a rounded stylus, a soft brush, and the wheel of a map-measuring instrument. The observer could identify the kind of instrument used without ever seeing it. He distinguished between brushing, rolling, and rubbing, and also between the blade, the point, and the rounded protuberance. There was always an objective component in the experience.

Friction at a single area of the skin. When a string is stretched over a region of the skin and then pulled in one or the other direction, the velocity is perceived. In order to study this phenomenon further, a yardstick sliding in a track was constructed on the table top, on which the observer could lower a fingertip, the hand being supported on an adjacent motionless block of wood. The yardstick could be moved laterally by a rack-and-pinion gear attached to a silent variable-speed clutch and a constant speed motor. When the observer lowered his finger

on this surface, he could feel the stick and report whether or not it was objectively moving, and the direction of the movement. He could estimate the speed and the amount of displacement over a given interval of time with some success. A speed of less than 1 mm. per second was detectable under these conditions. In this case there was no path of movement over the skin but only directional stretching and friction of a single patch of skin. There was no change in the "local sign" of stimulated points. Nevertheless the perception was that of a moving object.

Passive rotation of the joints of the limb. With the apparatus described, when the wooden support for the hand was removed and the fingertip was allowed to rest on the yardstick, the observer could still report accurately whether or not it moved, the direction of movement, its speed and the amount of displacement. Such results are predictable from those of Goldscheider (1898, described in Geldard, 1953, Ch. 12. See also James, 1890, Ch. 20). Only the receptors in the joints of the limb are stimulated. In the situation described, the joints of the finger, wrist, elbow, and perhaps the shoulder are all involved. The perception, as before, was that of a moving object. A naive observer would assert that he perceived the movement "by touch" although in actual fact the only cutaneous stimulation was a deformation of the skin of the finger, unchanged except by reason of slight tremor. The whole forearm was invisible behind the curtain.

Evidently three anatomically different receptive systems, two in the skin and one in the joints, can be functionally equivalent for the experience of objective motion when they are passively stimulated by an external agency.

Friction on the skin combined with active movement of the arm. In active touch, one runs the fingers over the protuberances of the object. In this operation friction on the skin and rotation of the joints are combined, together with voluntary contraction of the muscles. But in this case what one experiences is the movement of the hand, not an illusory motion of the object. We wished to make an even stronger test of this fact of phenomenal stability. Hence the apparatus was used to determine whether the motion of the yardstick, or its stability, could be detected while the finger was being actively moved along its surface, i.e., rubbed over it.

The fingertip of the observer was placed at one end of the stick and he was required to move his hand slowly either toward the right or the left. During this active movement the object was, on occasion, put into slow motion in the same or the opposite direction. It turned out that the existence, direction, and approximate speed of the object could be perceived, as before. Introspectively, there appeared to be a distinct separation of the movement of the hand from the motion of the stick. Both of these separate displacements were perceived relative to the space of the invisible table top hidden behind the curtain, this being connected to the visible and tangible floor and chair. One can only conclude that the propriospecific component in the total stimulus input had been separated from the exterospecific component.

Discussion. The ability of an individual to distinguish objective motion from subjective movement is a theoretical puzzle of long standing in visual perception (Gibson, 1954). A movement of the eyes is never confused with a motion of an object even though both involve what seems to be

the same retinal stimulus. Evidently the problem is analogous in tactual perception, although the solution can hardly be quite the same. When an individual actively moves his hand relative to his body, and thus relative to gravity and the surface of support, he is in contact with the earth as well as with the object, whether it be moving or motionless. A receptive system of higher order is excited during activity than the system during passivity. Hence the stimuli which have one excitatory capacity for a receptive subsystem in passive touch will have a different excitatory capacity in active touch—a different specificity. The relations between stimulation of the subsystems become significant. A lawful covariation of the inputs carries information different from that carried by isolation of the inputs, and yields different perceptions. Johannes Müller's law of the specific conscious quality of each excited nerve is evidently much too simple.

To lump one set of receptive systems together as *touch* and another as *kinesthesis*, then, is to obscure the function of the systems in combination. If both had their own unique quality of experience there would be some justification for the traditional belief in separate senses, but they do not. Anatomically different receptors may serve the same function and arouse the same experience, and, moreover, may serve different functions in different combinations. It might be better to drop the term sense and speak instead of "esthesic systems."

It should be noted that nothing has been asserted about *performatory* movements of the hand, as when an object is relocated or a tool is wielded. The perception is then subsidiary to action and the receptive systems take on still another function not here considered.

Perception of the Texture, Substance, and Shape of an Object by Touch

Katz (1925) showed that the texture of a solid substance could be perceived by touch if there was relative motion between the surface and the skin. Ordinarily this means rubbing the fingers over the surface; perpendicular motion was very much less effective than lateral motion. A substance could be identified by its texture; the observers could distinguish a dozen kinds of paper including blotting paper, wrapping paper, and writing paper. Katz also demonstrated the perceptibility of other physical properties of a surface besides rough-smooth, including soft-hard, and various kinds of plasticity, elasticity, or viscosity.

Katz was primarily interested in the fine structure of the surface of an object, and the substance of which it was composed. He had less to say about the *shape* of a substance, which is perhaps the main characteristic of an object. Active touch, however, is an excellent channel of spatial information in that the arrangement of surfaces is readily picked up. The solid geometry of things is best got by feeling them. A number of observations illustrate this assertion.

Curvature as against planarity of a surface. Departure of a surface from the plane can be judged by an observer either by holding his hand against the surface or by running his fingers over it. The opposite qualities of convexity and concavity are immediately given. On a smaller scale, protuberances or indentations on a plane surface are identified, and this acuity is taken advantage of in Braille. Moreover a surface with positive curvature on one axis and negative curvature on the other is perceived as saddle-shaped.

The accuracy of all these judgments could be but has not been measured.

Slant of a surface with respect to gravity. When the observer's hand is placed on an adjustable palm-board outside the field of view, it is found that he can judge its slant with precision. He can nullify this slant and adjust it to a normal plane. This holds for all three planes of space, the frontal, the saggital, and the horizontal. Variable and constant errors for settings to the normal plane seldom exceed 2°–3°. Furthermore, the slant of a surface that is visible only can be equated to the slant of a surface that is tangible only. The cross-modal matching of slant is accepted as natural by all observers, and the manual judgment can be used as a criterion of visual perception (Gibson, 1950).

When the palm-board is maintained at a slant from the frontal upright plane for some time, "normalization" seems to occur, and a negative aftereffect ensues. In this respect, the tangible slant of a surface behaves like the visible slant of a surface (Bergman & Gibson, 1959).

Parallelity of two surfaces. If a pair of adjustable palm-boards is held between the two hands, behind a curtain, they can be adjusted to the parallel position with small errors. The divergence or convergence of the two planes are felt as opposite qualities of a simple spatial experience. This dimension is also susceptible to normalization; after a few minutes of feeling divergence (say) the setting accepted as parallel is found to have shifted several angular degrees toward divergence. This fact has been studied in detail (Gibson & Backlund, 1963).

Distance between two surfaces. The span of the hands holding an object between them is experienced vividly. The tangible extent can be matched with a visible extent or vice versa. Likewise the distance between the thumb and one or more of the opposable fingers is perceived, and is accurately matched with the width of a visible object. According to Kelvin (1954) the cross-modal size comparison is made with no less accuracy than either unimodal size comparison alone. The finger-span of one hand can be immediately felt as equal or unequal to the finger-span of the other hand, and this comparison has been utilized by Köhler and Dinnerstein (1949) in studying a phenomenon they termed the kinesthetic figural aftereffect. In the many repetitions of this experiment the equivalence of the two hands for perceiving a spatial extent seems to have been taken for granted without its being studied in its own right.

Plane-angle, or edge of a surface. To exploratory touch, a junction of two planes constituting an edge is very noticeable. The angle is perceptible as varying from sharp to dull. At the extreme case of a knife-edge, sharpness is more accurately judged by touch than by vision.

Edge-angle, or corner of a surface. Where two edges intersect, or where three or more planes adjoin, a corner is produced. In the case when the angles of a convex corner are acute it is apt to be called a point. Corners and edges seem to be the main identifying feature of many objects. The fingers appear to seek them out and explore them.

Discussion. Planarity, curvature, slant, parallelity, span, edge, and corner may be conceived as variables of solid geometry. They are all perceptible to active touch, and appear to be simple qualities of experience. Since they are amenable to physical measurement, psychophysical experiments would be possible, but few have been performed.

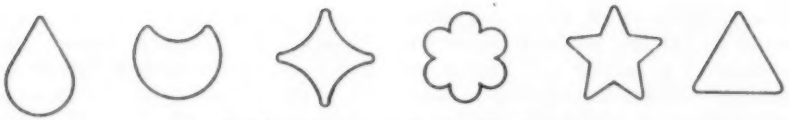


FIG. 1. The six forms to be identified by touch.

TACTUAL FORM PERCEPTION: A COMPARISON OF ACTIVE AND PASSIVE TOUCH

The study of the perception of objects by touch has been largely confined to investigations of two dimensional form or pattern on the skin, with passive stimulation (e.g., Zigler & Barrett, 1927). They are in the tradition of experiments on cutaneous localization and the two point threshold, the assumption being that the skin is first of all a receptor mosaic with patterns of excitation being projected to the brain. The forms pressed into the skin were called *solid* or *outline* depending on whether the edges of a surface or the edges of a metal strip were used; the latter gave better discrimination. The identification of forms in this manner is not very accurate, and there is some uncertainty even about circles, squares, and triangles.

Everyone knows, however, that a better way to perceive a thing of this sort is to run one's fingers over the edges. The fact is often taught to beginning students, and is even demonstrated in laboratory exercises (Langfeld & Allport, 1916) but a search of the experimental literature failed to locate its source. Lashley (1951), for example, asserts as if it were common knowledge that the same shape which can scarcely be distinguished with passive touch "can readily be distinguished when explored by tactile scanning" (p. 128). An experiment was performed to quantify this assertion, since no authority for it could be found. Each of a set of tactual forms had to be matched to its visual equivalent. The accuracy with which this could be done was compared when the form was pressed into the palm of the hand and when it was held above the palm and explored by the fingertips.

Method

The sources of stimulation were bent metal strips with handles, the edges having the forms shown in Figure 1. They were in fact small "cookie cutters" with a mean diameter of 1 inch. The six used were chosen for being mutually different in about the same degree. No two had the same number of corners or points. Each subject was shown the set of stimulators and the equivalent set of numbered drawings hung on the curtain in front of him. He then put his hand palm up on the table behind the curtain and was either touched by the object or was allowed to touch it for several seconds. In the one condition, he was instructed to open and relax his hand; in the other to cup the fingers upward and prepare to feel the edges. He then had to identify the form by number. The stimulator was not necessarily presented in the same orientation as the drawing.

The six forms were presented five times each (but the subject was told not to expect equal frequencies) and under two conditions, making a total of 60 trials. The order was random. No preliminary practice or knowledge of results was given. Twenty subjects were tested.

Results. A chance level of judgments would be 1/6 or 16.7%. For passive touch the mean frequency of correct matches was 49%. For active touch the mean frequency was 95%. The difference is significant.

In the active condition, the subject was permitted to explore the edges in any manner he chose. One style was to use all five fingertips simultaneously, another was to move the tip of the

index finger around the edge, and others were intermediate. All subjects, however, actively moved their fingers and thus obtained relative motions between the skin and the object.

In the passive condition, the form was pressed down on the palm by manual action of the experimenters so that a certain amount of unsteadiness was inevitable. When subsequently, with 20 different subjects, a mechanical lever system was used to apply the stimulation the mean frequency of correct matches fell to 29%.

Some subjects reported that they tried to count the number of corners or points they could feel, and that this was easier with the active than with the passive condition.

Discussion. A common belief has been verified but, in a sense, no claim can be made for the conclusiveness of the comparison. It could be argued that the fingertips are more "sensitive" than the palm of the hand, meaning that localization is better and the two-point threshold lower. This is true. The comparison is not very meaningful. The real point of the experiment is that tactual form perception in one meaning of that term does not depend on the pattern of local signs on the skin. If any number of fingertips from one to five can be used indifferently, the chaotic pattern of local signs resulting can hardly be the sole basis of the perception. With active touch no forms existed on the skin, but only a changing pattern of pressures. Evidently the registering of "form" by touch is not as straightforward a process as experimenters have supposed. Another experiment may help to clarify the matter.

TRANSPOSABILITY OF FORM ON THE SKIN

Gestalt theorists have emphasized the fact that a form can be moved over the retina, completely changing the set of receptors stimulated, without changing the perception. The contribution to perception of the anatomical pattern of excitation becomes a puzzle. The same is true of the skin.

Along with the experiment just described, another was carried out at the same time in order to compare passive static form perception with passive moving form perception. In these trials the "cookie cutter," instead of simply being pressed into the palm of the hand, was pressed and continually rotated clockwise and counterclockwise with a twisting motion. The same area of skin was stimulated in both cases and the same anatomical receptors were available, but in the latter case the form of stimulation changed from one instant to the next.

Judgments of the form of the object were more accurate when it was rotated than when it was not, the mean frequencies of correct matches being 72% and 49%, respectively. The difference is significant at the 2% level. To the writer's introspection, the corners of the object seemed to "stand out" better when it was twisted. In short, the form of the object seemed to become clear when the form of the skin-deformation was most unclear.

Discussion. Continuous change of the proximal stimulus with no change of the rigid object constituting the source of the stimulus is an old problem in visual perception. Why does the perception correspond to the form of the object instead of to the form of the stimulus? Evidently a quite similar problem arises for tactual perception. The paradox is even more striking, for tactual perception corresponds well to the form of the object when the stimulus is almost formless, and less well when the stimulus is a stable representation of the form of the object. A clear unchanging percep-

tion arises when the flow of sense impressions changes most.

It might be that the skin does not have as its primary function the registering of form as this has usually been conceived. Impressions of form and location relative to the skin might be quite separate from and incidental to the use of the extremities as sense organs. The informative stimuli might well be incorporated in the seemingly complex motions and transformations of the skin. There must exist relations of separation and proportion which remain invariant over time, and which are specific to the object. They do not represent the shape of the object but they specify it. The solution to the problem of object perception in touch would then be that continuous change in the proximal stimulation is accompanied by nonchange, that is, the set of invariant relations. The former is not noticed; the latter is separated out and attended to.

The role of exploratory finger movements in active touch would then be to isolate the invariants, that is, to discover the exterospecific component in the flux of stimulation. Only thus could the above paradox be resolved.

Relation between Touch and Vision

A number of similarities and differences have been noted in the course of these observations between active touch and visual perception. Many of these have been pointed out before by other observers such as Katz (1925, 1935) but a few of them have not previously been emphasized.

Surface perception. Solid surfaces are tangible and visible. This can only be because they have mechanical inertia in the first place, and because they reflect light with scattering in the second place. (A few visible surfaces such as smoke are not tangible and a few tangible surfaces such as plate glass may not be visible, but most surfaces are both.) The physical texture of a surface, the various kinds of roughness, can be observed either by vision or by touch. The color of a surface is visible only, not tangible, and consequently the pigment-texture of a surface (the differential coloring of marble or of a picture) is the same. On the other hand, if color is intangible temperature is invisible. Each sense has its special sensitivities to the properties of a surface but they also have some in common. The bark of a tree looks rough without touching it and feels rough without seeing it. The expert can identify the tree by either sense alone.

The visible array of surfaces extends for miles of course, and the tangible array only to the length of man's arm. But if surfaces determine space perception, both touch and vision are spatial senses.

Object perception. A detached solid substance with a topologically closed surface is perceptible as such to both the hands and the eyes. By virtue of physics, the eyes can encompass the closed contour of a very large object (depending on its distance) while the hands can explore only the surface of an object of limited span. The unity of the visual perception is said to be based on the "figure-ground" phenomenon, the simultaneous registering of the whole contour, whereas the unity of the tactual perception has to be based on either cutaneously separated impressions or on successive impressions. Vision is said to register only the front surface of an object while touch registers the front and the back at the same time. In these respects vision and touch seem to differ. Nevertheless these contrasts are exaggerated; succession enters into

the operation of both senses. The eyes normally fixate in succession just as the fingers explore in succession. With two eyes, and by changing one's standpoint, more of an object than its front surface is perceived. As regards the unity of separated impressions from the two hands, there is a visual analogy in the unity of separated impressions from the two retinas.

Spatial properties of objects. Certain variables of the shape of a detached object, slant, curvature, edge and corner are independently visible and tangible. The equivalence of the two modes of perception for judgments of the object is such that differences got by one sense are equated to differences got by the other. The stimulus information for these cross-modal judgments, one depending on mechanical and the other on luminous energy, might seem hopelessly discrepant, but the fact is that the judgments can be made.

Whether both vision and touch can register form is an ambiguous question, as the experiments showed. The answer depends on what is meant by form. Drawings and pictures on a flat surface are sensed only by vision. The conformation of the edges of an object, however, is sensed by both vision and active touch. The tangible and visible features of things abstracted by solid geometry (slant, curvature, edge, corner) did not seem any more complex to introspection than the solely visible features of things abstracted by plane geometry (triangle, square, circle).

A new set of shapes has now been devised for the study of object-perception by active touch. They consist of ten solid sculptures, or free-forms, made of plastic, the surfaces being curved, with no planes, edges, or corners. They are intended to be felt with two hands (and are called "feelies"). Approximately one-half of each surface (the "rear") is convex; the other half (the "front") consists of six convexities with intermediate saddles or concavities. In general, there are five protuberances around a central protuberance, but no object is symmetrical, either radially or bilaterally. They cannot, therefore, be distinguished from one another by counting. Each is readily discriminated from every other by vision of the "front" surface. They are also mutually distinguishable by feeling, although with some error and hesitation for an unpracticed observer.

Replicas of the ten objects are available, made from the same molds. It is therefore possible to present one object to the hands and the same or a different object to the eyes simultaneously. Preliminary results indicate that the cross-modal matching of these novel preceptions is possible even for a naive observer, and that practice can bring about errorless judgments in all observers so far tested.

Passivity and activity in touch and vision. In passive touch the individual makes no voluntary movements. Similarly, in passive vision he makes no eye movements, which means that he must voluntarily fixate his eyes on a point specified by the experimenter. Neither state is natural to an individual. In a tactual situation, the observer will explore with his fingers unless prevented and, in a visual situation, he will explore the focussable light, fixating, accommodating, converging and pursuing. Both senses are normally active. The passive stimulation of the skin or the retina is necessary for the study of the receptor-cells in the skin or the retina, but the experiences resulting are atypical.

In active touching and looking the observer reports experiences of a quite different order. They correspond to

the environment instead of to the events at the sensory surface. The experiences noted with passive stimulation can scarcely be noticed, if at all. The effective stimuli are presumably different, being relations and combinations selected from a flux of potential stimuli. Moreover, the experiences of active touching and looking are more nearly equivalent; they are more alike to introspection than those of passive touch and vision.

In general, the foregoing survey suggests that vision and touch have nothing in common *only when they are conceived as channels for pure and meaningless sensory data.* When they are conceived instead as channels for information-pickup, having active and exploratory sense organs, they have much in common. In some respects they seem to register the *same* information and to yield the *same* phenomenal experiences.

CONCLUSIONS AND SUMMARY

A series of observations, both introspective and behavioral, confirms the distinction between *touching* and *being touched.* The former is a channel for a great variety of information about the environment, but whether it should be considered one or several senses is a matter of definition. The simple formula that it consists of passive touch plus kinesthesis is insufficient. The hypothesis of two components of stimulation, one exterospecific and one propriospecific, is more promising.

Many properties of surfaces and objects can be registered by active touch. The accuracy of such judgments can be measured and formal experiments are needed. The observations suggest that the variables of solid geometry can be more profitably studied than those of form. Object constancy is characteristic of tactual perception. In several respects, tactual and visual perception are alike.

In general, experimenters have not realized that to *apply* a stimulus to an observer is not the same as for an observer to *obtain* a stimulus. Obtainable stimuli can be controlled and systematically varied, just as applied stimuli can be. Psychophysical experiments are possible for tactual perception as well as for cutaneous sensitivity if appropriate methods are devised for controlling the obtainable stimuli.

REFERENCES

BERGMAN, R., & GIBSON, J. J. The negative after-effect of the perception of a surface slanted in the third dimension. *Amer. J. Psychol.*, 1959, **72**, 364–374.

BORING, E. G. *Sensation and perception in the history of experimental psychology.* New York: Appleton-Century, 1942.

GELDARD, F. A. *The human senses.* New York: Wiley, 1953.

GELDARD, F. A. Adventures in tactile literacy. *Amer. Psychologist*, 1957, **12**, 115–124.

GIBSON, J. J. The perception of visual surfaces. *Amer. J. Psychol.*, 1950, **43**, 367–384.

GIBSON, J. J. The visual perception of objective motion and subjective movement. *Psychol. Rev.*, 1954, **61**, 304–314.

GIBSON, J. J., & BACKLUND, F. An after-effect in haptic space perception. *Quart. J. exp. Psychol.*, 1963 (in press).

GOLDSCHEIDER, A. *Gesammelte Abhandlungen von A. Goldscheider.* Vol. II. *Ueber die Bewegungsempfingung.* Leipzig: Barth, 1898.

JAMES, W. *Principles of psychology.* New York: Holt, 1890.

KATZ, D. *Der Aufbau der Tastwelt.* Leipzig: Barth, 1925.

KATZ, D. *The world of colour.* (Trans. by R. B. MacLeod & C. W. Fox) London: Kegan Paul, 1935.

KELVIN, R. P. Discrimination of size by sight and touch. *Quart. J. exp. Psychol.*, 1954, **6**, 23–34.

KÖHLER, W., & DINNERSTEIN, D. Figural after-effects in kinaesthesis. In, *Miscellanea psychologica Albert Michotte.* Louvain: Editions de l'Institut Superieur de Philosophie, 1949. Pp. 196–220.

LANGFELD, H. S., & ALLPORT, F. H. *An elementary laboratory course in psychology.* Boston: Houghton Mifflin, 1916.

LASHLEY, K. S. The problem of serial behavior. In L. A. Jeffress (Ed.), *Cerebral mechanisms in behavior.* New York: Wiley, 1951, Pp. 112–136.

NAFE, J. P., & WAGONER, K. S. The nature of pressure adaptation. *J. gen. Psychol.*, 1941, **25**, 323–351.

REVESZ, G. *Psychology and art of the blind.* New York: Longmans Green, 1950.

ROSE, J. E., & MOUNTCASTLE, V. B. Touch and kinaesthesis. In *Handbook of physiology.* Washington, D. C.: American Physiological Society, 1959. Ch. 27.

ZIGLER, M. J., & BARRETT, R. A further contribution to the tactual perception of form. *J. exp. Psychol.*, 1927, **10**, 184–192.

(Received July 30, 1961)

Psychological Review
1962, Vol. 69, No. 6, 492–500

A SYSTEM FOR SHAPE RECOGNITION

J. A. DEUTSCH [1]

Department of Psychiatry and Psychology, Stanford University

In this paper it is suggested how certain morphological features of the histology of the visual system may on simple assumptions function to discriminate distance, angle, and shape. For the purposes of this paper shapes can be regarded as a derivative of distance between points on the contours and also as a derivative of the angles between parts of the contour. It is difficult to know whether in the human case it is the relative distance cues or the cues from angles which are more relied upon. In the case of a visually more primitive organism, such as the rat, though it is known that it can discriminate distance and also relative size, its abilities to discriminate the angle of one line to another or the curvature of lines remains untested. It is therefore entirely possible that such animals lack a mechanism to enable them to discriminate angles. If this is the case the shape discriminatory abilities of such organisms are based mainly on comparisons of relative distance between contours.

It would seem likely that in discriminating shape animals utilize prior information available in the nervous system about distances between contours and about angles between contours also. However, the basis of the capacity to estimate distance between any two points in the visual field or the angle between any two contours in the visual field is obscure. If we wish mechanically to assess the relation between two things or events such as the distance between two points, it is necessary (though not sufficient) that these two things or events jointly determine some third event.

On the face of it, a system of fixed connections would appear to be very uneconomical in the number of structural elements which would have to be employed to assess distance (Sutherland, 1959). Suggestions have therefore been made to decrease the number of elements required to relate visual points or contours so that they can jointly determine some third event. One such way is to transform distance in space into distance in time by some kind of moving process, a notion familiar from television. One such suggestion of scanning from fixed coordinates external to the shape to be analyzed has come from Dodwell (1957) for the rat and Deutsch (1960) for the octopus. Deutsch (1955, 1958, 1960b) has suggested that such a conversion of distance into time is effected not by a scan external to the shape to be analyzed but by a wave of excitation propagated by the contours of the shape itself.

However, further investigation of the problem has led the writer to the conclusion that a system relying on a conversion of spatial separation into temporal interval is not necessary. For instance an examination of the neurohistological data concerning the zona plexiforme of the deep retina in the octopus has led him to postulate a system for shape recognition which does not rely on these principles (Deutsch, 1960a), though no functional fixity of

[1] The writer is indebted to John Gilbert for help in computing the "outputs" of the suggested system. He also wishes to thank G. A. Horridge and T. H. Bullock for valuable discussions.

connections has been postulated. There is good correspondence between this system and what can be seen of the striking morphology of the zona plexiforme and the properties of the shape discriminating system of the octopus as they have been ascertained so far. However, the shape discriminatory capacities of the octopus are different from those found among the vertebrates. It seems nevertheless that even in other cases besides the octopus it is possible to devise a system of reasonable economy in which the principle of transforming distance into time is not necessary.

Specifications for a Biological Shape Recognizing Mechanism

In a previous paper on the subject (Deutsch, 1955) the present author enumerated some of the peculiarities of shape recognition in animals which a theory should explain. These peculiarities are not common to all animals, such as the octopus, and here special theories are necessary (such as Deutsch, 1960), but these peculiarities appear to be those of the general case of higher vertebrates. (Unfortunately the experimental evidence in this field leaves much to be desired. But if it is argued that the properties set down below have not been established experimentally to be the case, then we can regard the properties listed below as predictions instead from the theory from which they are deducible.)

These are some of the abilities that a system must have before it can be considered a theory of shape recognition.

1. It must be able to produce the same output independent of the location of the image on the receptor surface. Shapes retain their identity in spite of changes in their position in the visual field.

2. The system must produce the same output independent of the rotation of the shape on the receptor surface. (For some of the evidence on this point see Gellermann, 1933, and Deutsch, 1960b).

3. A similar output should be given for the same geometrical shape of different sizes, but in such a way that transfer to a larger shape is more efficient than transfer to a smaller shape.

4. Mirror images should produce an identical output. They cause great difficulty in discrimination in animals and children and are readily confused even by adults.

It should at this stage be emphasized that only a partial set of visual classifying mechanisms is being suggested. The fact that we can perform visual discrimination of the tilt, position, or brightness does not constitute a disproof of a theory of distance, angle, and shape recognition. It is clear that the mechanism which classifies such properties per se must be discarding information about tilt, position, or size. It must be an abstracting mechanism par excellence. The fact that an animal can visually discriminate other properties besides these simply means that it is also possessed of other visual mechanisms. Therefore, if an animal can be shown to discriminate between two stimuli this does not disprove the notion that a particular system is not to be found inside the animal; it could be supplemented by another mechanism. On the other hand an inability to discriminate, which would be predicted by the theory, would tend to support it as would a tendency to transfer to other stimuli, identical on the theory, even where discrimination can be taught.

Let us begin by examining a neurohistological picture (Figure 1) of the bee's optic lobes (from Cajal & Sanchez, 1915) which gives a structure

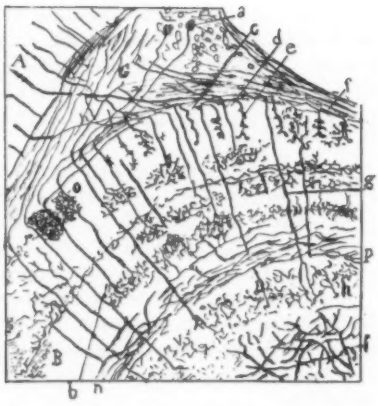


FIG. 1. Neurohistological picture of the bee's optic lobes.

well suited for the transformations to be postulated.

As all the authors have recognised, the plexiform mass of this lobe appears to be concentrically furrowed, by means of concentric bands, of a finely granulated aspect, which appears to Radl to constitute the anatomical characteristic of the visual centres of the animal series, and which remind not a little of the plexiform layers of the optic lobes of the fishes and batrachians, so minutely described by P. Ramon (pp. 99–100, author's translation).

Such a structure is not unique to the bee, though the particular section here is unusually good for the purpose of illustration. It is to be observed that the lobe is divided by various horizontal layers, three of which are indicated (at g, p, and i). The retina is not in the section, but fibers from it are shown proceeding from the direction of A (upper left-hand corner) which enter the lobe at right angles to the horizontal layers after crossing at the chiasma. It can be seen that in the first layer to be reached by the retinal fibers the number of terminations is large, and the size of the terminal arborizations is small by comparison with those further inside the lobe. From this it is possible to infer that in this layer is preserved information about the activity of a larger number of ommatidia than in layers further in and also that the ommatidia which are represented in this layer are closer neighbors to each other spatially in the eye than those represented by fibers which terminate on subsequent layers. The fibers proceeding from the retina entering further layers are fewer, presumably from points further apart on the retina, and their terminal arborizations are larger than in the previous layers, thus giving scope for interaction over a wider angle. So it seems that as we go inwards from the retina, progressively fewer fibers cross each horizontal layer and even though the fibers are converging (because of the concentric arrangement of the layers) the extent of the terminal arborization at each level progressively increases.

Given such an anatomical arrangement, we need to make some assumptions about how the parts behave in order to speculate about the functions of this arrangement. The only assumption which is necessary is that the fibers in each horizontal layer act as spatial summators. We need only postulate that if two incoming retinal fibers which are spatially contiguous are also simultaneously excited, then this will excite fibers in the horizontal layer. The sum of the excitation generated in this manner in each layer will reflect the number of neighboring fibers that are simultaneously active.

As the fibers which are neighbors in the various layers are drawn from angular distances progressively further apart, there will be distinctive patterns of activity emerging from the horizontal layers depending on the overall distribution of widths and breadths of visual areas. A totally uniform excitation of all the visual ele-

ments will cause maximum activity in all the horizontal layers. A visual field composed of contrasting areas each of which spans that angle of field from which one receptor is selected to end in a particular horizontal layer will produce a minimum of activity in this layer. In such a case, alternate fibers ending in this layer will be inactive, and so spatial summation will not occur. As we inspect the output from horizontal layers which sample closer neighborhoods, the rate of activity in these will increase as we move away from the layer in which activity is at a minimum. The same will occur as we move in the other direction where fibers drawn from more distant points end in horizontal layers further down.

The system so far described will perform a kind of crude autocorrellogram of an area with itself, indicating the number of times areas of a given size repeat regardless of direction. Each horizontal layer performs this task for an area of one approximate size. Therefore, a uniform excitation of the whole visual field should produce a maximum of excitation on all layers. On the other hand, a checkerboard with squares of a certain size should cause great lowering of excitation in one particular horizontal layer, as the repetition frequency (taking all directions together) of an excited area will of necessity be small. It can be seen that the system described for the bee could not perform the type of discriminations to be seen in the rat, for instance. However, we have supposed, in order to be economical in our assumptions, that it samples receptors from gradually increasing distances at each layer. It is evident that such a system would perform very poorly when confronted with outline shapes and best when dealing with regularly repeated blobs or stripes. This property tallies with what is known of the bee's visual abilities. It is not supposed that the vertebrate system is of precisely the same type. The example of the bee is quoted in order to show that the kind of structure necessary to perform our postulated transformation is a biological possibility. This structure can without much difficulty be assigned the kind of task which we require.

The higher vertebrates appear to take contours as the basis of their shape discrimination. Therefore, neighboring arborizations of fibers at successive layers should be made to send a pulse to the horizontal layers only when both of them are active and a fiber should only be made active if it has been stimulated by a contour discontinuity. Further, fibers descending to progressively lower layers cannot merely be samples of retinal stimulation and progressively larger distances, but must give information about the presence of a contour at within these distances. Given these alterations, we have a system which is capable of discriminating distances. A particular distance between two retinal points will be represented as activity in a certain horizontal layer. As a result it can be used as the basis for a certain kind of system of shape discrimination.

System. The basic function the postulated system performs is to measure the distance between all pairs of points on the boundary of the shape and so produce a distribution of how many pairs of points there are at each distance from each other. This may be made clearer by an example. If we take four points at the corners of a square, we can then measure the distances between all the pairs of points. The histogram representing this figure will show a total of four at the distance equal to one of the sides of the square and a total of two at the distance equal to the length of the diagonals. We can perform a similar transformation of a

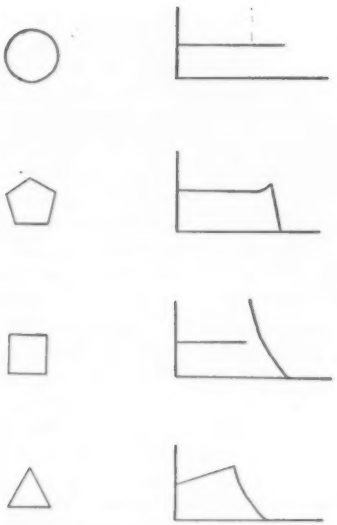


FIG. 2. Histogram measuring the distance between all the pairs of points.

shape given any number of points on its boundary to derive the curves shown in the illustration (Figure 2). As can be seen, different shapes under such a transformation still remain different from each other.

Given that any two points of a certain distance from each other will cause activity on one particular horizontal layer, then if there is more than one pair of points at that distance from each other, more activity on that one particular horizontal layer will result. Given a set of points at varying distances from each other, activity on successive horizontal layers will show the distribution of these distances. By inspecting the output of these layers we have the information of how many points there are at the distance from each other, at the distance +1, and so on. We thus have a system performing the type of transformation postulated above and which also conforms to all the specifications (except Number 3) laid down above. It should be clear that the first specification has been obeyed. The same output will be produced by the horizontal fibers independent of the position of the vertical fibers impinging on the horizontal layer with their arborizations, given that channels equal distances apart retinally end on the same horizontal layer. Similarly, rotation of the optical input will not produce any change unless we postulate special and regular distortions of the distribution of the receptor surface or special collecting properties of the horizontal layers. Finally, mirror images will produce the same output. The output is simply a measure of the distances between the points in a figure and these are the same in mirror images.

The specification which has not been fulfilled is that relating to the invariance of the same shape in spite of different size. This requires some further considerations. We have seen that in the optic lobe of the bee that the number of fibers relative to the distance along the retina decreases at each successive layer as the fibers travel from the retina. In this way the output from the horizontal layers which would be sending signals out in succession as two lines of fixed length were moved farther and farther apart would decrease layer by layer. Now the successive thinning out of the fibers could be such that the firing from these layers would not decrease as the lines were increasing in length in proportion to their distance. If we descend layer by layer (measuring the number of neighbors at progressively larger distances) the output generated by a small shape will be the same as the last portion of the output generated by the same larger shape. In this way the pattern for a small shape will be completely reproduced if we transfer the animal to a discrimination to larger shapes, though it will be added to at

the beginning, whereas if we transfer to smaller shapes, the larger shape to which the animal was originally trained will have the first part missing. The problem of how the same relative message appearing on different horizontal layers is then taken and reacted to as the same message is still of course unsolved. There are various alternatives, but it is probably better to leave this problem by saying that the third specification looks as if it could be fulfilled. Shape would be given by the relative position of the signals on the horizontal layers to each other, size by their absolute position on these layers.

A remark which should be made about the system as it stands is that any coincidences (signaling that two contour components are within a certain distance) will be repeated at each succeeding horizontal layer, though they will not be signaled at levels which measure smaller distances. There are ways of removing this redundant information. For instance, we can alter our system in such a way that all the fibers from the retina are taken through all the layers with increasing overlap of arborization at each level. We can then set the system in such a way that only the boundary or end of an arborization and the center of another arborization (that is another fiber) will produce a response in the horizontal layer when they are both on. Though there will not be a reduction of the number of fibers traveling down to each succeeding level in this case, we may still suppose that the area over which fibers interact to cause a signal in the horizontal layer increases as we descend downwards.

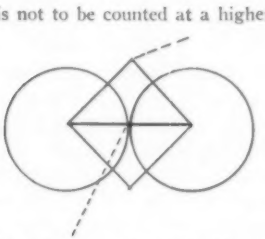
Another Mechanism

The mechanism which was derived from the neurohistological picture of the bee's optic lobes is not the only kind of system which would produce the postulated transformation and satisfy the stated specifications. The suggestion may be worth making as neurohistologically distinct variants of the same principle could have developed.

This second mechanism would also generate the same signals, given the same shape, irrespective of its locus on the receptor surface and its tilt on the frontal parallel plane. Further, it will confuse mirror images in any orientation (as a young child will). Also, this mechanism will reduce shapes of different size to a similar message in the same way as the mechanism described above. That is, the message for a small shape will be the same as one end of the message signaling the same but larger shape. This would be consistent with the well known asymmetry of transfer between smaller and larger shapes. It is easier to obtain transfer to a larger shape from a smaller than vice versa. When transferring to a larger shape the animal will obtain the complete message it has learned together with an additional portion. However, when transferring to a smaller shape, some of the message representing the shape it has learned will be missing. We might therefore reasonably assume that transfer would be worse in this second condition.

In the mechanism to be considered, let us assume that there is a three-dimensional space filled with cells. Each cell on the top of this solid sends down connections into the solid. Each cell or point is conceived as radiating a large number of connections in all directions, but still at a constant angle. Each point at the top may thus be conceived as the apex of a cone, and the connections as lines on the surface of this cone converging at the apex of the top of the solid. (We can, for convenience, think of these as lines regularly spaced on the surface of the cone.) When a contour is projected on

the top of this multicellular solid, each cell at the base will send down excitation into its set of connections fanning-out downwards as on the surface of a cone. If we picture such ascending excitation as black lines on the surface of each cone, these lines on two different cones will intersect at a height depending on the distance of the apices of these cylinders. To place a limitation on the number of intersections of lines between the two cylinders to two, we shall arbitrarily assume that only head-on collisions between lines of excitation will be registered as coincidences. That is, only lines at a 180-degree angle to each other will be counted as intersecting or, in other words, lines intersecting at the tangent of the circumference of the two cones.



In this way the distance of two points at the top of the solid will determine the height at which the excitations emanating from each will coincide. The number of coincidences at a particular height or horizontal layer will be determined by the number of points (or rather cells) at a particular distance from each other on the contour of a particular shape. Thus the arrangement outlined above will produce a histogram or a spectrum for each particular shape, with distances along one ordinate and the number of points at each distance from each other on the other ordinate. The spectrum or histogram produced will be the same independent of the position of the shape at the base of the solid and also independent of its orientation. Mirror images will also give rise to the same spectrum. The property of generalization to the same shape of different sizes is yet to be accounted for. Here it is possible to have recourse to the same kind of principle as in the first type of mechanism outlined above. We need merely suppose that the number of possible coincidences decreases by the factor of the distance downwards. For instance, there may be a certain critical distance between two ascending lines as they pass near each other within which a coincidence signal will be generated if both lines are excited. As the lines pass downwards on the cone they will be progressively more widely spaced on the surface of the cone and therefore increasingly less likely to be within the critical distance of each other, even if their angle with regard to each other is sufficiently correct. If, furthermore, the coincidences are counted by layers progressively further apart (in terms of retinal distance) the histogram generated by a small shape will be the same as the end of the time series of the same large shape.

The advantage of this second scheme is that it could be made to yield good discriminations between various degrees of curvature. This would occur if we assume that once coincidence had occurred between two paths of descending excitation, the excitation would not descend any further. In this way the excitation generated by a straight line would all coincide on the first layer and there would be no further coincidences whereas an increase in curvature would bring about further coincidences (Figure 3).

Discrimination of Angle

So far two systems have been described which yield discrimination of

distance between points. Points are brought into relation at some third point and this third point varies along a single continuum depending on the distance between the two points. Further, all the pairs of points with the same distance between them are represented by activity in the same structure. This structure is conceived of as a layer, so that a particular distance between any two points is translated into activity in a particular layer. It was shown that shape recognition could be based on such a system. However, as was stated above, discrimination of size of angle between two contours is also a capacity of the visual nervous system. It will be shown briefly below how the type of mechanism already outlined can simply be expanded to provide this capacity.

It will be apparent that the rate at which two contours converge or diverge will be reflected in the angle at which successive coincidences between points occur in the horizontal layers. For instance, two parallel lines will cause a set of coincidences which is horizontal to the successive layers. An increase in angle between lines will cause a gradual tilting away from the horizontal of the successive points of coincidence. Such information could be read out from the system in different ways. One such possibility would be to have cells with fields of arborizations shaped like discs or flattened spheres sending a signal out whenever two coincidences occurred within the field. With considerable overlap of these fields, angle would be measured by the number of cells sending a signal out. If this information was sent out layer by layer, a mechanism analogous to the one based on distance could operate on the basis of angle. Such functioning of cells and their arborizations is not implausible in view of the work of Maturana, Lettvin, McCulloch, and Pitts (1960) on the frog's visual system.

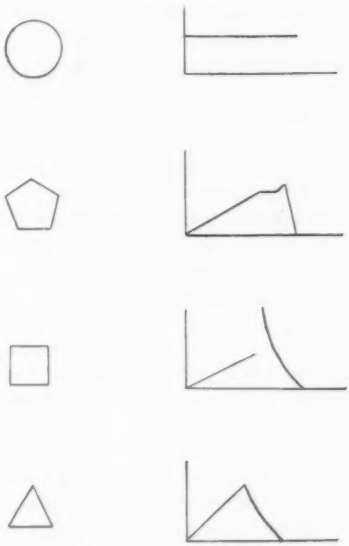


Fig. 3. Histograms produced by second mechanism.

Discussion

The question of uniqueness of transformation of a given shape has not been so far dealt with. No mathematical proof is so far forthcoming that shapes or patterns other than left-right or up-down mirror images are uniquely transformed. However, extensive tests of a trial-and-error nature have failed to reveal different shapes with identical profiles, though shapes that appear similar give rise to profiles with certain common features.

The "preset" systems described above give rise to certain determinate principles of classification, and so of similarity. Such principles seem to operate in the case of our own shape recognition and therefore they were set up as specifications for a theoretical system. The similarities were therefore built into the system, in con-

trast with such attempts as those of Uttley (1956) and Taylor (1956). Here equivalences must be learned. Unless the machine has some already built in classification systems to enable it to abstract the principles on which it is being made to equate diverse groups of dots, its economy is bound to be low as compared with a preset system, which already abstracts the same properties as the human operator. Its performance in novel situations is also likely to be poor. Sutherland (1959) has made a strong case for supposing that there is a preset classifying mechanism. If there is, it is important from the point of view of mechanical shape recognition to find out what it is. The ideal machine would classify shapes in the way we do and so would see similarity in distorted human output, such as handwriting, where distortion is permitted because it seems that the essential elements of identity are not lost.

REFERENCES

CAJAL, S. R., & SANCHEZ, D. *Contribucion al conocimiento de los centros nervosos de los insectos.* Vol. 13. Madrid: Universitad de Madrid, Trabajos del Laboratoria de Investigaciones, 1915.

DEUTSCH, J. A. A theory of shape recognition. *Brit. J. Psychol.*, 1955, **46**, 30–37.

DEUTSCH, J. A. Shape recognition: A reply to Dodwell. *Brit. J. Psychol.*, 1958, **49**, 70–71.

DEUTSCH, J. A. The plexiform zone and shape recognition in the octopus. *Nature, London*, 1960, **185**, 443–446. (a)

DEUTSCH, J. A. *The structural basis of behavior.* Univer. Chicago Press, 1960. (b)

DODWELL, P. C. Shape recognition in rats. *Brit. J. Psychol.*, 1957, **48**, 221–229.

GELLERMANN, L. W. Form discrimination in chimpanzees and two-year-old children. I. Form (triangularity) *per se*. *J. genet. Psychol.*, 1935, **42**, 3–27.

MATURANA, H. R., LETTVIN, J. Y., MCCULLOCH, W. S., & PITTS, W. H. Anatomy and physiology of vision in the frog (*Rana pipiens*). *J. gen. Physiol.*, 1960, **43**, 129–176.

SUTHERLAND, N. S. Stimulus analysing mechanisms. In, *Symposium on the mechanization of thought processes.* London: Her Majesty's Stationery Office, 1959.

TAYLOR, W. K. Electrical simulation of some nervous system functional activities. In C. Cherry (Ed.), *Third London Symposium on information theory.* London: Butterworth, 1956. Pp. 314–327.

UTTLEY, A. M. Temporal and spatial patterns in a conditioned probability machine. In C. E. Shannon & J. McCarthy (Eds.), *Automata studies.* Princeton: Princeton Univer. Press, 1956. Pp. 277–285.

(Received March 27, 1961)

Psychological Review
1962, Vol. 69, No. 6, 501–519

PATTERNS OF ADRENERGIC-CHOLINERGIC IMBALANCE IN THE FUNCTIONAL PSYCHOSES

LEONARD S. RUBIN

Eastern Pennsylvania Psychiatric Institute, Philadelphia

During this last decade remarkable strides have been made in the treatment and control of the so-called functional psychoses with the aid of chemical agents. In view of the current status of the science of chemotherapy in this area, it is surprising that such beneficent results have been reported because there is no widely accepted chemical theory of the functional psychoses. A therapist is unable to prescribe, on the basis of preliminary laboratory examinations (of bodily fluids, for example), the chemical agent that is relevant to any specific endogenous chemical disequilibrium. One empirical chemotherapeutic approach consists of administering drugs on the basis of phenomenological impressions of simple behavioral dimensions such as depression or excitation. It need not be emphasized that few empirical findings and no theoretical formulations extant specifically relate these behavioral dimensions to any disturbed tissue system or state of chemical disequilibrium that is reversible by chemical intervention. In another approach, the tacit assumption is made by therapists administering drugs that the population of functionally psychotic individuals is homogeneous. In both approaches, when improvement does occur in some fraction of the population, it had not been predicted.

Prior to the formulation of several theories that have served to stimulate a great deal of recent research, numerous empirical biochemical and physiological studies were performed in the hope of uncovering some characteristic endogenous feature that differentiated normals from psychotics. The following observations are representative of a larger number of studies that suggest that the functional psychoses have as a concomitant some disorder of homeostasis. Fischer (1953) reported that schizophrenic blood was toxic to tadpoles, while Federoff and Hoffer (1956) reported it toxic to fibroblasts grown in tissue culture. The latter researchers, however, found the blood of patients undergoing surgery equally toxic. This suggested that a nonspecific perseverative response to stressful stimuli may be involved. In addition, various measurements of the response to stress have been made employing the following variables which reflect autonomic activity: pulse rate, blood pressure, skin circulation, and response to adrenalin (Finkelman & Stephens, 1936; Pfister, 1938). The results of these studies on chronic schizophrenics generally indicate that they are less reactive than healthy people. When a large series of measurements was made of basal metabolic rate, circulation rate, body temperature, excretion of 17-ketosteroids, thyroid activity, and water metabolism in schizophrenics and in healthy people, it was found that the variance of the measurements was greater for the schizophrenics (Richter, 1957). These studies again suggest that a concomitant of schizophrenia is some disorder of homeostasis.

Devoid of specific deductively elaborated hypotheses, this plethora of empirical findings contributed by physiologists, pharmacologists, and behav-

ioral scientists stimulated the need for a rational approach to chemotherapy. Several astute attempts at systematization have been made. They will be described briefly because they serve to lay the groundwork for the major thesis contained in this report; that is, that discrete patterns of autonomic dysfunction may be etiological factors in the functional psychoses.

The major contemporary chemical theories of the functional psychoses are inextricably dependent upon the behavioral and biochemical effects of two psychotomimetic agents, mescaline and lysergic acid diethylamide (LSD-25) (Rubin, 1957). A biochemical theory of schizophrenia put forward by Osmond and Smythies (1952) and stated more specifically by Hoffer, Osmond, and Smythies (1954) postulated the existence of an endogenous mescaline-like substance, the "M substance," a toxic metabolite formed by the defective metabolism of epinephrine. They claimed that, whereas normal individuals metabolize epinephrine by amine oxidase or sulfoesterase, the psychotic metabolizes epinephrine by a phenolase. The resulting production of quinone indoles, i.e., adrenochrome and adrenolutine, interferes with cerebral metabolism. Hoffer (1957) administered adrenolutine (a more stable isomer of adrenochrome) and a placebo to a group of subjects and he attempted to determine which persons had been given adrenolutine and which the placebo. The results of the experiment were equivocal. Gastaldi, as reported in a review by Smythies (1958), evaluated the effect of the more highly reactive, unstable adrenochrome and reported the absence of behavioral abnormalities. In a recent study (Holland, Cohen, Goldenberg, Sha, & Leifer, 1958), equilibrium plasma concentrations of adrenaline and noradrenaline were measured in healthy people and schizophrenics. No significant difference was found in the rate of utilization of circulating adrenaline and noradrenaline. Axelrod (1958) attempted to detect the presence of adrenochrome in schizophrenics and found that it was absent in schizophrenics as well as in healthy people. As yet, this chemical theory of schizophrenia that postulates the existence of an endogenous metabolite of adrenaline produced by a deficiency in the process of methylation lacks empirical verification.

The above biochemical theory of schizophrenia was established following the observation that mescaline, an hallucinogen, was similar in chemical structure to the neurohumors, adrenaline and noradrenaline. The second theory to be considered was formulated after the observation was made that the most powerful hallucinogen, LSD, was a potent peripheral antagonist of serotonin (5-hydroxytryptamine). The clue to the important role of serotonin came from its discovery in the brain. Like acetylcholine, the substance is present in nervous tissue in a precursor state and is active only in an unbound form. The amine is unevenly distributed in the brain, its concentration being highest in the brainstem, especially the hypothalamus, lowest in the cortex, and almost undetectable in the cerebellum. The high biological activity of this amine strongly suggests an important function in chemical mediation in subcortical centers. Additional support for the role of serotonin in neural transmission comes from consideration of the distribution of the enzyme monoamine oxidase, the enzyme that destroys serotonin, and 5-hydroxytryptophan decarboxylase, the enzyme that synthesizes serotonin. Both enzymes are found in highest concentration in the hypothalamus. These considerations led Brodie and Shore (1957) to

hypothesize that serotonin was an important chemical mediator of neural activity in the brain. Although the hypothesis has been fruitful, no current evidence has established beyond doubt that serotonin plays a transmitter role or is auxiliary to the action of other transmitters. Woolley and Shaw (1954) accepted the presumptive evidence for serotonin as a central neural transmitter substance and went on to postulate that the hallucinogenic effect of LSD might be due to its interference with the action of serotonin in the brain. According to these authors, the action of LSD, as well as other alkaloids, can be ascribed to the indole moiety in the structures of LSD and serotonin. Consequently, compounds containing the indole ring may act as antimetabolites to serotonin. This suggested to them that mental aberrations were produced by endogenously produced indole compounds that competed for serotonin receptor sites in the brain. This attractive theory has been challenged by at least two recent observations. Cerletti and Rothlin (1955) found that a close analogue of LSD, 2-bromo-D-lysegic acid diethylamide, was as effective as LSD in antagonizing several peripheral actions of serotonin, yet this antagonist of serotonin was found to be devoid of hallucinogenic action in humans. It should be noted that chlorpromazine is also a potent antiserotonin and, instead of being hallucinogenic, is a powerful antihallucinogen employed in the treatment of psychotics.

As indicated by the foregoing review, considerable effort is being expended to understand the biochemistry of both the LSD and the mescaline psychotic-like states in normal individuals. Although the evidence provided by researchers who have worked within these conceptual frameworks has been inconclusive, the quest for the specific toxic endogenous metabolite goes on. However, we still quite realistically speak of the chemical control of schizophrenia and not the chemical treatment.

A more detailed critical review of these theories led the author (Rubin, 1959) to conclude that it was far from ineluctable that they would ultimately lead to an effective chemotherapy for the functional psychoses. At this point, the author would introduce his theory by stating that the functional psychoses are characterized by patterns of *disordered* neurohumoral activity in the central representation of the autonomic nervous system. Since adjustments of the internal milieu are mediated by the autonomic nervous system and integrated in the hypothalamus, it is necessary to consider the humoral nature of hypothalamic activity before the theory is elaborated and supporting evidence is presented. Specifically, the assumption will be examined that the central representations of the autonomic nervous system are cholinergic for the parasympathetic division and adrenergic for the sympathetic division.

Vogt (1954) examined the concentration of sympathin (a combination of norepinephrine and epinephrine) in the different parts of the central nervous system of the dog and found that sympathin, like acetylcholine, possessed a specific pattern of distribution. Two regions of the brain, the hypothalamus and the area postrema, stood out above all others by their high concentrations of norepinephrine. This fact suggests, though it does not prove, that these amines play a part in the specialized function of those regions of the brain in which their concentration is high. Vogt's conclusion regarding the correlation between sympathin content and central representation of sympathetic functions was summarized in the following statement:

With exception of the area postrema, those regions which contain the highest concentration of sympathin—the hypothalamus, parts of the midbrain and the floor of the fourth ventricle—also contain the central representation of sympathetic activity (p. 473).

However, when Vogt attempted to localize the site of sympathin within the hypothalamus, it was found that the anterior and posterior hypothalamus did not show consistent differences in sympathin content. This is of some interest because of the controversy over whether or not the anterior hypothalamus is the site of parasympathetic, and the posterior hypothalamus is the site of sympathetic centers. In this particular histochemical aspect, the two regions resemble each other. When the hypothalamus was divided into dorsal and ventral parts, however, a definite concentration gradient was found. The ventral region contained more noradrenaline than the dorsal region. Based upon Vogt's findings, the position taken in this paper is that the central representation of sympathetic function in the hypothalamus contains the highest concentration of the biogenic amines; norepinephrine, epinephrine, and serotonin.

In attempting to ascertain whether the central representation of the autonomic nervous system may be cholinergic for the parasympatheic division, the following facts must be considered. For purposes of description, the hypothalamus in man may conveniently be subdivided into four regions: the supraoptic middle region located above the optic chiasm and rostral to it, the tuberal or infundibular middle region, the mammilary middle region, and the lateral region. For a purpose that will become clearer as the discussion continues, the composition of the supraoptic middle region will be described. It includes the nuclei supraopticus, paraventricularis, suprachiasmaticus, supraopticus diffusus, and the anterior hypothalamic area (Kuntz, 1953). Feldberg and Vogt (1948) attempted to map out, in the central nervous system of the dog, the distribution of the enzyme (or enzyme system) that synthesizes acetylcholine. They found a very high value of enzymatic activity in the region of the supraoptic nuclei. By their definition, this portion of the hypothalamic area may be considered cholinergic. However, the high figures for the supraoptic nucleus do not suggest that cholinergic fibers originate from cells in this region, rather, such fibers terminate in synaptic connection with these cells. Unfortunately, the enzymatic activity of the anterior hypothalamus, the controversial site of parasympathetic activity, was not determined.

It seems reasonable to conclude from the foregoing discussion that specific areas within the hypothalamus cannot be viewed as exclusively adrenergic or cholinergic. The evidence does suggest that the hypothalamus contains large quantities of the biogenic amines as well as components of the cholinergic system. Consequently, in the remainder of this paper, when describing the central adrenergic-cholinergic imbalance that characterizes the functional psychoses, I do not necessarily implicate specific sympathetic and parasympathetic central structures as such.

Since Cannon's (1932) formulation of his Emergency Theory of Emotion, it has been accepted generally that, whenever an organism is confronted with a sudden emergency, the resulting massive autonomic response is functionally adaptive. Specifically, a massive and complete mobilization is achieved by the sympathetic division of the autonomic nervous system with the resultant secretion of great quantities of epinephrine. As Grenell states in Fulton (1949), "The so-called 'fight,

flight or fright reaction' is a product of maximal generalized sympathetic activity." One major feature of an organism's response to an emergency situation, therefore, is a significant increase in the release of certain sympathomimetic amines. The ascription of a role to the parasympathetic division of the autonomic nervous system during the emergency reaction has been conspicuously neglected. Its role has been confined to the protection, conservation, and restoration of bodily resources. The hypothesis here offered is that the requisite *maximal* generalized adrenergic activity ensuing the perception of threat is facilitated by a decrease in cholinergic outflow. If we consider, as Cannon did, the functional utility of the increased adrenergic response to stress, then it follows that maximal mobilization can be most effectively attained under a condition in which central cholinergic outflow actually decreases below levels that characterize the organism during nonstressful steady states. It is conjectured that, were cholinergic outflow to increase or remain at the level of nonstressful states during the time the organism is mobilizing for an emergency, the net adrenergic response would be less than maximal. As the effectiveness of the organism's adjustment to emergency is directly related to maximal mobilization of bodily resources, it is hypothesized that maximal mobilization can only result from increased activity of central adrenergic centers and reduction of central cholinergic outflow. As a corollary, it is inferred that the patterns of autonomic dysfunction characterizing the functional psychoses consist of significant departures from the characteristic interactions of the adrenergic and cholinergic components of the normal individual at rest or under stress.

It is further postulated that neurohumoral imbalance in the functional psychoses may involve either or both neurohumoral systems and that pathological variation in either system is not unidirectional. Conceivably, either malfunction of the adrenergic mechanism or the cholinergic or both may lead to significant departures from that reaction requisite for maximal mobilization of the organism. A hypothetical matrix of varieties of neurohumoral imbalance has been constructed by reviewing the efficacy and mode of action of certain drugs already employed with varying success in the treatment and control of the functional psychoses. Patterns of cholinergic and adrenergic malfunction in the psychoses may thus be inferred from pharmacological and clinical observations. From the following evidence, it has been deduced that neurohumoral dysfunction may be characterized by either an excessive or deficient adrenergic response mechanism.

A Disordered Adrenergic Response Mechanism in the Functional Psychoses

Iproniazid (Marsilid), an amine oxidase antagonist that inhibits the oxidative deamination of epinephrine, has been used extensively in the treatment of mental depression. *In vivo* studies on animals with iproniazid indicate that it can produce a relatively rapid and large (threefold) rise in brain serotonin (Zeller, Barsky, & Berman, 1955). Before proceeding to the statement of an important inference derived from these observations, it would be well to evaluate the controversial role of monoamine oxidase in the metabolism of biogenic amines.

There are at least four potentially possible metabolic routes for the enzymatic conversions of norepinephrine, epinephrine, and closely related catecholamines: deamination by mono-

amine oxidase, oxidation to adrenochrome and related products by cytochrome oxidase and tyrosine oxidase, conjugation to the glucuronide or sulfate, or O-methylation by catechol-O-methyl transferase (Koelle, 1959). Though the importance of monoamine oxidase in the oxidation of catecholamines is controversial, it is well established that dopamine, a precursor of norepinephrine and epinephrine, is oxidized by monoamine oxidase at a reasonably rapid velocity. Furthermore, according to Bacq (1949), there is little possibility that the catecholamines are oxidized to indole derivatives by cytochrome oxidase *in vivo*. As far as the third metabolic pathway listed above is concerned, the results of a great deal of work suggest that free catcholamines and their immediate metabolites from other pathways are conjugated in considerable proportions. Finally, LaBrosse, Axelrod, and Kety (1958) have suggested that catechol-O-methyl transferase is the enzyme of major importance in the metabolism of circulating norepinephrine and epinephrine. They arrived at this conclusion after they observed that over 90% of a 1-mg. dose of dl—epinephrine given to normal human subjects by intravenous infusion—was recovered as the free or conjugated form of the primary product, metanephrine, or its subsequent oxidation product, 3-methoxy-4-hydroxymandelic acid. It is important to note that the authors still invoke a role for monoamine oxidase. It participates in the oxidation of metanephrine to the mandelic acid derivative. Pekkarinen, Castrin, and Israel (1958) present evidence suggesting that amine oxidase acts directly upon the catecholamines. This position is supported by the work of Shore (1958) who found increased amounts of norepinephrine in the brain and other organs of small animals after treatment with iproniazid. On the basis of his studies, he concluded that: "norepinephrine, as well as serotonin, is normally metabolized in vivo largely through the action of monoamine oxidase." He criticized much of the experimental evidence which suggested that monoamine oxidase is unimportant in catecholamine metabolism by emphasizing the erroneous use by others of *in vitro* tissue preparations. When monoamine oxidase was blocked *in vivo* by Marsilid, Shore could not find any alternate major pathway for the metabolism of norepinephrine. Evidence for or against each metabolic route cannot begin to be adequately described in this paper. A reasonable conclusion is that while monoamine oxidase may not be the sole enzyme system involved in the degradation of the catecholamines, it must be considered a significant part of the mechanism.

To return to the effects of iproniazid on the concentration of biogenic amines, it is likely, according to Truitt (1958), that this drug, through its inhibition of monoamine oxidase in the brain, potentiates the action therein of endogenous stimulant biogenic amines (possibly norepinephrine, epinephrine, and serotonin). The side effects (constipation, difficulty in micturition, hyperreflexia, sweating, and dryness of the mouth) are also compatible with a hypothesis of peripheral sympathetic stimulation. The implication drawn from these observations is that *some* psychotics who are benefited by iproniazid are defective in some brain catecholamines, or in central adrenergic stimulation.

When one turns to the treatment or control of the schizophrenic reaction, it is found that reserpine and chlorpromazine have been extensively used. Hyperemotional responses mediated by sympathetic centers in the

hypothalamus are markedly depressed by the action of reserpine. Peripherally, the syndrome produced by reserpine is marked by bradycardia, slight lowering of blood pressure, stimulation of peristalsis, and gastric hypersecretion. Brodie and Shore (1957) proposed that reserpine prevented serotonin binding and thereby caused constant activation of the vegetative, parasympathetic centers. In addition, reserpine was found to deplete the store of biogenic catecholamines in the body as determined by the increased excretion of 5-hydroxy-indoleacetic acid, a metabolite of serotonin. Evidence has been presented in Truitt which favors a role for serotonin as a central excitatory agent at least in subcortical centers. The hyperactivity of serotonin, rather than its deficiency, therefore, may play a dominant role in mediating uncontrolled emotional responses in acute psychoses. Thus, the mechanism of reserpine action may be by depletion of brain concentration of serotonin and other catecholamines rather than through the sedative action of released serotonin. These observations therefore suggest that *some* functionally psychotic patients for whom reserpine therapy is efficacious may be characterized by an *excessive* adrenergic activity in subcortical centers. This same conclusion is also applicable when chlorpromazine is considered, for, in addition to its peripheral antihistaminic and anticholinergic effects, it has been shown to have an adrenergic blocking action.

In summary, it is postulated that the neurohumoral dysfunction that characterizes *some* psychotics may be due to either (*a*) a neurohumoral deficiency of some sympathomimetic amine or to some enzyme disorder such as monoamine oxidase overactivity in the adrenergic excitatory mechanism of the brain; or, (*b*) a neurohumoral overabundance of some sympathomimetic amine or to some enzyme disorder such as monoamine oxidase underactivity in the adrenergic mechanism of the brain.

A Disordered Cholinergic Response Mechanism in the Functional Psychoses

There are suggestions that the cholinergic component of generalized autonomic activity is also an important parameter of the functional psychoses. Forrer (1956) has recommended the use of large doses of tertiary atropine in the treatment of manic-depressives in the manic phase. He claimed that the patients in the manic phase of manic-depressive psychosis responded remarkably well and immediately to this chemotherapy but he offered no mechanism which could possibly account for the effectiveness of this drug. As the pharmacological action of atropine is that of a cholinergic blocking agent, it may be inferred that *excessive* brain acetylcholine or deficient acetylcholinesterase with manifest excessive cholinergic outflow is important to the understanding of the chemical fundament in some psychotics. A strong theoretical statement which in part supports this inferred mechanism as to the efficacy of atropine therapy has been offered by Hoffer et al. (1954). He has claimed that excessive parasympathetic activity, with concomitant increases in the concentration of acetylcholine, is an essential feature of the schizophrenic process. In contrast to Forrer, however, Hoffer believes that cholinergic blocking agents as well as cholinergic compounds should aggravate the schizophrenic process and perhaps make psychotics worse.

Paradoxically, Pfeiffer, Jenney, Gallagher, Smith, Bevan, Killam, Killam, and Blackmore (1957) have reported

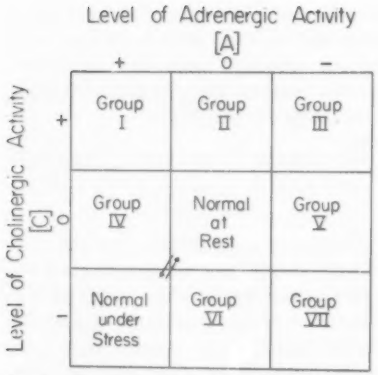


FIG. 1. Patterns of adrenergic-cholinergic interaction for normals and psychotics.

that some schizophrenics respond with a lucid interval to stimulation by cholinergic stimulants such as pilocarpine, arecoline, and 2-dimethylamine ethanol. The implication of this report is that some psychotics are characterized by *defective* cholinergic responsiveness. Recently, Rubin (1958a, 1958b) has reported that a functionally psychotic sample of patients was significantly differentiable from normals by virtue of the rate of hydrolysis of acetylcholine chloride by erythrocyte cholinesterase. All of the patients employed in the study manifested a defect in rate of hydrolysis. Some subjects hydrolyzed the substrate too rapidly and others too slowly. The inference made, on the basis of the kinetics of the enzymatic reaction, was that some psychotics may be deficient in cholinergic activity while others may be characterized by excessive cholinergic activity.

PATTERNS OF DISORDERED ADRENERGIC-CHOLINERGIC STEADY STATES IN PSYCHOTIC DISORDERS

A hurried review of the evidence presented seems replete with paradoxical facts and theories. At first it appears that an effective chemotherapy for some psychotics should counteract deficient adrenergic activity. Then, it seems that more effective drugs which decrease adrenergic activity should be developed. Additional conflict and confusion is raised by assertions of efficacy for cholinergic as well as anticholinergic drugs. These apparent contradictions may be resolved by regarding the population of functional psychotics as consisting of a specifiable number of discrete patterns of disordered adrenergic-cholinergic levels of reactivity. From the preceding review of the pharmacological action of certain drugs on the neurohumoral components of the autonomic nervous system, it is assumed that in the population of psychotics there may exist three levels of adrenergic activity: overreactivity, deficiency, and a normal level from which directional deviations are measurable. The cholinergic outflow may also be described as consisting of three levels: overreactivity or deficiency of cholinergic levels as measured by reference to a level characteristic of the unstressed normal. The matrix of adrenergic-cholinergic steady states derivable from such a paradigm is presented in Figure 1. In this figure, A represents central adrenergic activity and C central cholinergic activity. The levels of activity of each dimension are symbolized by +, 0 and −. The positive sign represents excessive activity, the zero, normal activity, and the negative sign, deficient activity. The combination of AoCo represents the levels of adrenergic and cholinergic activity that characterize the normal human adult at rest. The eight other combinations represent hypothetical adrenergic-cholinergic levels of interaction. Previously, the hypothesis had been developed that a normal, ordered response to stress consisted of maximal generalized sympathetic activity that was

attained when cholinergic stimulation was decreased (C−), while adrenergic stimulation continued to increase (A+) up to that maximal level which characterized the individual. This characteristic adjustment is contained in Figure 1, and may be represented as A+C−. It is now further hypothesized, within the framework of the general theory, that the remaining seven interactions between adrenergic-cholinergic activity represent the discrete patterns which comprise the population of functional psychotics. To recapitulate, the major hypotheses elaborated from the above considerations are: (*a*) Effective autonomic adjustment to emergency situations requires a *maximal* generalized adrenergic response; (*b*) To attain a maximal level of sympathetic activity, it is not only necessary that increased levels of adrenergic activity follow the emergency but that the level of cholinergic activity decrease; (*c*) The functional psychoses are characterized by patterns of neurohumoral dysfunction that deviate significantly from the pattern of the normal, healthy individual at rest, as well as from the adrenergic increase and cholinergic decrease that characterizes the normal individual during stress; and (*d*) The varieties of neurohumoral dysfunction represent discrete levels of activity of interacting adrenergic and cholinergic mechanisms. That is, each pattern of neurohumoral dysfunction represents an impairment of function of either the adrenergic or cholinergic mechanism or both.

PUPILLARY REACTIVITY AS A METHOD FOR INVESTIGATING PATTERNS OF DISORDERED ADRENERGIC-CHOLINERGIC STEADY STATES IN PSYCHOTIC DISORDERS

The hypotheses elucidated would remain interesting speculations unless a method were described which could systematically measure the relevant parameters. To this end a physiological measure of adrenergic and cholinergic activity was developed for investigating the patterns of disordered autonomic steady states in psychotic disorders. The miotic response of the pupil to light (constriction) and the mydriasis ensuing the absence of light (dilatation) were employed to measure cholinergic and adrenergic mechanisms, respectively. The rationale of the procedure is suggested by the following considerations. The cholinergic nature of the activity of the oculomotor nerve under the influence of light has been proved by the positive assay for acetylcholine in the anterior chamber of the eye (Engelhart, 1931) and by the disappearance of the light reflex after the injection of cholinesterase (Mendel & Hawkins, 1943). Conversely, the anticholinesterases induce miosis, presumably by interfering with the hydrolysis of acetylcholine by acetylcholinesterase. The adrenergic nature of the dilatory response of the pupil is suggested by the work of Poos (1939), Hess and Koella (1950), and Hess, Koella, & Szabo (1950) on isolated radial preparations of the iris. These studies demonstrated that the iris muscle contracted in response to small concentrations of adrenalin and noradrenalin. Since these substances are responsible for neurohumoral transmission in sympathetically innervated structures, they believed it reasonable to assume that a dilator muscle existed that responded to sympathetic excitation. In addition, Frohlich and Loewi (1910) had shown that the mydriasis induced by adrenalin is intensified by cocaine, a monoamine oxidase inhibitor.

Since pupillary constriction had been shown to be dependent upon the liberation of acetylcholine, and dilatation

upon an adrenergic mediator, the rate of dilatation was considered to represent the kinetics of a general adrenergic mechanism, while the rate of constriction was considered indicative of the kinetics of the cholinergic mechanism. If the varieties of disordered neurohumoral function that characterize the functional psychoses do represent discrete levels of activity of interacting cholinergic and adrenergic mechanisms that differ significantly from the patterns manifested by normals at rest and under stress, then different patterns of constriction and dilatation should be observable. That is, it was assumed that peripheral cholinergic and adrenergic imbalances, respectively, mirror central cholinergic and adrenergic imbalances.

PROCEDURE

Although the apparatus and procedure for obtaining measurements of pupillary dilatation and constriction have been described elsewhere (Rubin, 1960, 1961a, 1961b), it seems appropriate to emphasize certain features of the procedure for increased clarity in exposition. To obtain the rate of dilatation for each subject, the following procedure was employed. After the subject was seated with his head positioned properly in a head rest, the room light was extinguished. After 5 seconds in the dark, an intense adapting light of 38.6 ft.-C. was focused on the subject's right eye for 10 seconds. At the termination of this interval, a photograph of the maximally constricted pupil was taken automatically. This was done twice in an effort to equate for previous conditions of adaptation of the retina. Each subsequent photograph of the pupil was taken after the subject had been dark-adapted for a prescribed interval of time following maximal constriction. The intervals of dark-adaptation employed were 5, 10, 15, 30, 45, and 60 seconds. The procedure employed to obtain measurements of the rate of pupillary constriction was initiated soon after measurements of pupillary dilatation had been obtained. Each photograph of the pupil was taken after the subject had been light-adapted for a specific interval of time following maximal dilatation. The intervals of light-adaptation employed were 1, 2, 3, 4, and 5 seconds. Prior to obtaining a measurement of constriction at each of these intervals, the eye was light-adapted for 10 seconds at a constant intensity before it was permitted to dilate maximally within the 60-second dark-adaptation period.

PATTERN OF ADRENERGIC-CHOLINERGIC ACTIVITY OF THE NORMAL ADULT UNDER STRESS

To test the hypothesis that normal adult subjects react to emergencies with an increased adrenergic response and decreased cholinergic response, measurements of pupillary dilatation and constriction were obtained on six normal, healthy adults between the ages of 19–40 (Rubin, 1960). Due to the variety of perceptual responses that are elicited in humans to the same stimulus, it is difficult to define a stimulus situation that is realizable in a laboratory as an "emergency" or as a "threat." Consequently, following the control measurements obtained when the subjects were at rest, each subject was required to immerse the right hand and forearm in a cold water bath maintained at 5° C. while measurements were made of dilatation and constriction. When the rate of pupillary dilatation at rest was compared to the rate obtained during the stress condition, it was found that the slopes of the curves were equal (parallel) but that the absolute level of dilatation under the cold pressor condition was significantly greater (Figure 2). Assuming that the responsiveness of the dilator pupillae is an increasing monotonic function of the amount of adrenergic mediator liberated at the effector site, it may be concluded that under stress a greater quantity of the adrenergic mediator is liberated. Therefore, the finding that the cold pressor test increased pupillary dilatation significantly was interpreted as supporting

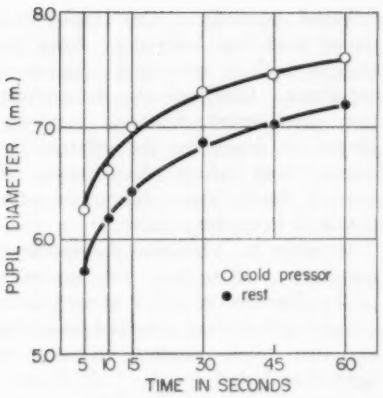


FIG. 2. Pupillary dilatation in normal subjects exposed to the cold pressor test.

the hypothesis that under a condition of stress, the normal human adult responds with an increased adrenergic outflow.

When the rates of pupillary constriction over a period of 5 seconds for the normal individuals at rest and under the condition of the cold pressor were compared, a significant difference was obtained (Figure 3). As opposed to the relationship observed during dilatation, the level of constriction of the pupil under the condition of the cold pressor test was decreased. This suggested a decreased cholinergic outflow in normal humans under stress, when it was assumed that the degree of constriction of the sphincter pupillae was an increasing monotonic function of the amount of acetylcholine liberated at the effector sites.

The significance of these findings, however, is contingent upon two basic assumptions: that the cold pressor test is representative or equivalent to other forms of physiological or psychological stress, and that the adrenergic and cholinergic reactivity of the pupil is representative of adrenergic-cholinergic mechanisms of interaction in vital neural integrative centers. If these assumptions are accepted, the following interpretation of the results of the experiment may be made with reasonable assurance. Upon exposure to a stress-engendering situation, the normal human organism responds with an increased adrenergic outflow and a decrement in cholinergic outflow. This reciprocal balance between the systems insures maximal mobilization of the resources of the organism required for "fight, flight, or fright."

ADRENERGIC-CHOLINERGIC REACTIVITY LEVELS OF FUNCTIONAL PSYCHOTICS

Several additional experiments were performed by the author (Rubin, 1961a, 1961b) in order to test two additional hypotheses. The first hypothesis tested was that the population of psychotics exhibited three distinct levels of adrenergic activity and three distinct levels of cholinergic activity. The second hypothesis tested was that psychotics manifest patterns of neuro-

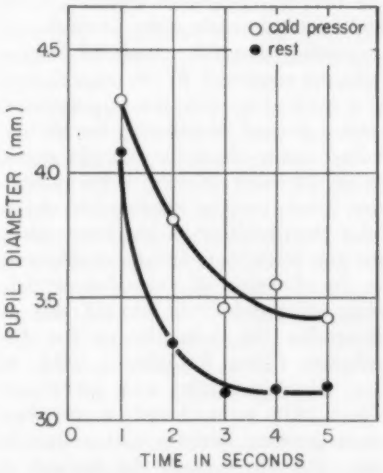


FIG. 3. Pupillary constriction in normal subjects exposed to the cold pressor test.

humoral dysfunction that deviate significantly from the pattern of normals at rest, as well as from the adrenergic increase and cholinergic decrease that characterizes the normal person during stress. Furthermore, the varieties of neurohumoral dysfunction could be represented by discrete patterns of levels of adrenergic-cholinergic interaction, as described previously in Figure 1.

To test these hypotheses, measurements of dilatation and constriction were obtained on 23 healthy, control subjects between the ages of 19–40 years. To determine the extent of dilatation between the fifth and sixtieth seconds, independent of the individual differences in the final level of dilatation attained, the following measurement was employed. The initial level of dilatation (mm.) was subtracted from the final level (mm.) and the difference was divided by the final level of dilatation. The ratio obtained in this manner for each subject represented the increase in diameter of the dark adapted pupil as a function of the maximal dilatation attained at the end of 60 seconds spent in darkness. Assuming that the sample of healthy subjects employed in the experiments is a random sample from a population with a normal distribution for the obtained ratio, then the sample mean, as an unbiased estimate of the population mean, may be employed to determine the likelihood of obtaining values for the ratio that deviate significantly in the direction of excessive or deficient dilatation. A graphic test to determine the normality of the distribution (Bliss & Calhoun, 1954) of the dilatation ratio was performed. The rankits were plotted against their corresponding variates and a straight line was fitted easily by inspection. This strongly suggested that the sample of dilatation ratio scores was distributed normally. One could now make inductive inferences from the sample with a reasonable degree of assurance. Consequently, the cumulative distribution function was employed to determine the relative frequency with which observations fell beyond the extreme dilatation ratios obtained from the sample.

In order to determine the extent of pupillary constriction, measurements of the diameter of the 1 minute dark-adapted pupil were obtained following exposure to a constant intensity of light that lasted for 1, 2, 3, 4, or 5 seconds. A constriction ratio for each subject was obtained in the following manner. To avoid negative values, the final level of constriction (mm.) was subtracted from the invariably greater initial level (mm.) and then divided by the final level. The graphic test for normality of the distribution of the ratios was performed and it was found that they were distributed normally. The cumulative distribution function was again employed to determine the likelihood of obtaining constriction ratios that fell beyond the extreme constriction values of the sample.

The probability of obtaining dilatation or constriction ratios significantly greater than or less than the extreme values in each distribution was then employed to evaluate the measurements obtained from a sample of psychotics.

The results to be described were obtained on two samples of patients who were tested at two different psychiatric institutions. As the results of the first experiment w. e confirmed by the second experiment, the following description pertains to 71 patients who had been diagnosed as functionally psychotic (Rubin, 1961b). Employing the probability limits obtained for the distribution of the dilatation

ratio in the sample of normal individuals, it was found that 21 of 71 patients manifested excessive dilatation, 26 of 71 patients manifested deficient dilatation, and 24 of 71 patients manifested normal dilatation. These relationships are presented in Figure 4 where the mean dilatation ratio for each of the three groups at each time interval is presented as a function of the amount of time spent in darkness.

Inspection of the figure clearly indicates that there are three levels of dilatation that characterize the sample of patients employed in the experiment. The hypothesized three distinct levels of adrenergic activity, as measured by pupillary reactivity, seem to have been confirmed. That is, if it is assumed that the magnitude of pupillary dilatation is proportional to the quantity of adrenergic mediator liberated, then it may be concluded that some of the patients liberated a quantity that was characteristic of normals at rest; others liberated more of the adrenergic mediator or had some enzyme disorder such as monoamine oxidase underactivity; and some liberated too little of the adrenergic mediator or had some enzyme disorder such as overactivity of monoamine oxidase.

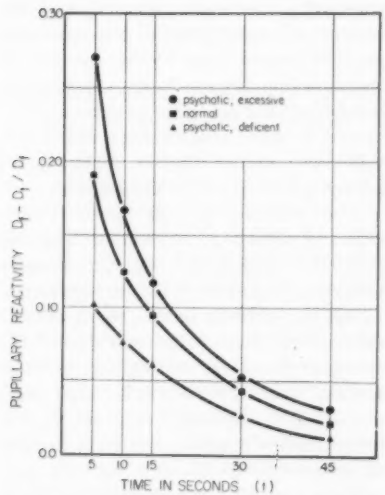


FIG. 4. Reactivity levels of pupillary dilatation in psychotic patients. (The values of the ordinate, D_f–D_t/D_t, consist of the final level of dilatation, mm., after 60 seconds of dark adaption, D_f, minus the level of dilatation at each interval in darkness, D_t, divided by the final level of dilatation, D_t. The abscissa represents the intervals of darkness in seconds, t, preceding photographing of the pupil.)

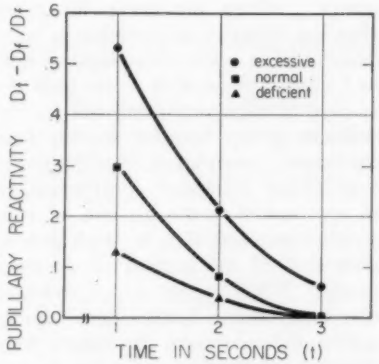


FIG. 5. Reactivity levels of pupillary constriction in psychotic patients. (The values on the ordinate, D_t–D_t/D_t, consist of the average diameter—mm.—of the constricted pupil at each interval, D_t, minus the diameter of the maximally constricted pupil after 5 seconds of light adaptation, D_t, divided by D_t. The abscissa represents the duration of light adaptation in seconds, t.)

Employing the probability limits obtained from the distribution of the constriction ratio in the sample of normal individuals, it was found that 36 of 71 patients manifested excessive constriction, 7 of 71 patients manifested deficient constriction, and 28 of 71 patients showed normal constriction. These relationships are presented in Figure 5 where the average constriction ratio for each of the three groups at each interval is represented as a function of length of time of exposure to a constant intensity light

source. When the curve for those showing defective constriction is compared to the curve representing normal constriction, it is seen that the amount of change in pupil size for the deficient group remains smaller than for normal constriction. For the group manifesting excessive constriction, it is apparent that the amount of constriction per unit time is much greater than that of the normal or defective group. With respect to constriction, therefore, three levels, each significantly different from the other, were found in the sample of patients employed in the studies. Assuming that the tendency toward constriction of the sphincter pupillae is proportional to the quantity of acetylcholine liberated, the demonstration of three levels of constriction suggests that: (*a*) in one group of patients, either excessive quantities of acetylcholine were liberated or there was deficient acetylcholinesterase activity; (*b*) in a second group of patients, insufficient quantities of the neurohumor were liberated or enzyme activity was excessive; (*c*) the third group manifested that level of cholinergic activity that was equivalent to the level of normal subjects. These observations on the kinetics of adrenergic and cholinergic activity confirmed two hypotheses deduced from the theory; that is, the existence of three distinct levels of adrenergic activity and three distinct levels of cholinergic activity.

In order to establish that there are varieties of neurohumoral dysfunction that comprise the population of functional psychotics, it remained necessary to show that the disorders consisted of an impairment of function of either the adrenergic or cholinergic mechanism, or both. The demonstration of these patterns of interaction was derived from the following considerations. Twenty-four out of 71 psychotic patients, it will be recalled, showed no impairment in the extent of dilatation. However, *all* of the members in this group manifested *aberrant* constriction. Twenty of the 24 manifested excessive constriction, while 4 patients were deficient in constriction. Conversely, of the 28 psychotic patients out of 71 who showed no impairment in constriction, *all* were aberrant in the extent of dilatation; 12 showed excessive dilatation and 16 manifested deficient dilatation. In confirmation of the major hypothesis, it may be stated that all of the 71 patients employed in the studies showed a significant impairment in either the tendency to dilatation or to constriction or both. Furthermore, if the magnitude of dilatation is considered to be proportional to the quantity of adrenergic mediator liberated, and the amount of constriction of the sphincter pupillae proportional to the quantity of acetylcholine liberated, then it may be concluded that all of the psychotics employed in these studies were differentiable into seven discrete patterns of adrenergic-cholinergic imbalance.

The varieties of adrengeric-cholinergic dysfunction elucidated indicate that the population of functionally psychotic individuals is heterogeneous. Thus, the current quests for a chemical panacea that affects only one of the components of the autonomic nervous system undirectionally do not seem promising. Instead, the results of these studies suggest that each pattern of adrenergic-cholinergic activity may require a specific combination of drugs which differentially affect the level of activity of either or both components (Table 1). The value of the present study is that it seems to provide a rationale for the selection of patients who require distinct chemotherapies as determined by their specific pattern of adrenergic-cholinergic imbalance.

An obvious but unwarranted implication that may be drawn by the reader of the foregoing assertion is that the author believes with warrantable assertability that these patterns of autonomic dysfunction are etiological factors in the functional psychoses. Admittedly, it is difficult to conceive of an environmental process that could produce central neurohumoral dysfunction. However, the experiments described in this report were not designed to answer the question, "Is central neurohumoral dysfunction the invariable antecedent of psychotic behavior or the effect of dynamic, environmental variables?" Although the answer to the question may be of academic interest, it would be more fruitful to accept the attitude of a pragmatist and insist that truth is preeminently to be tested by the practical consequences of belief.

TABLE 1

A CLASSIFICATION OF FUNCTIONAL PSYCHOSES INTO DISCRETE LEVELS OF ADRENERGIC-CHOLINERGIC INTERACTION AND THE DERIVED CHEMOTHERAPIES

Group	Adrenergic-Cholinergic reactivity pattern	Suggested chemotherapy
I	A+C+	Adrenolytic agent in addition to an anticholinergic compound
II	AoC+	Anticholinergic compound
III	A−C+	Adrenergic stimulant in addition to an anticholinergic compound
IV	A+Co	Adrenolytic agent
V	A−Co	Adrenergic stimulant
VI	AoC−	Cholinergic stimulant
VII	A−C−	Adrenergic stimulant in addition to cholinergic stimulant

ADRENERGIC-CHOLINERGIC DYSFUNCTION IN PSYCHOTIC CHILDREN

The technique of employing pupillary measurements to determine adrenergic-cholinergic steady-states that characterize adult psychotics was also applied to ascertain whether autistic, psychotic children were significantly differentiable from normal children. An additional comparison could readily be made between those steady-states that characterized normal children and the adrenergic-cholinergic responsivity of normal adults. Thus an opportunity was also afforded to contribute to an understanding of the ontology of autonomic reactivity in normal individuals.

In the preliminary experiment to be described, data on dilatation and constriction was obtained on four normal, healthy children between the ages of 7–9.5 years, and five autistic children between the ages of 7–12 years. A statistical analysis of the rate of dilatation between these groups indicated that the slopes were equal and parallel but separated significantly by a constant (Figure 6). Again, assuming that the magnitude of the response of the dilator pupillae is an increasing monotonic function of the amount of adrenergic mediator liberated at the effector site, then the absolute level of adrenergic activity for the normal children is greater than that for the psychotic children. This suggests that normal children have a significantly greater capacity for increased adrenergic outflow than autistic children. When the two groups of children were compared with respect to the rate of constriction, it was found that the rate was significantly faster for the normal children, although no significant differences obtained in the final, absolute level of constriction (Figure 7). This result was interpreted as indicating that the rate of cholinergic activity was more rapid for normal

children. These results suggest that normal children have a greater capacity for sympathicoadrenal response, and, of equal importance, a greater manifest capacity to reestablish more quickly the vegetative functions mediated by the cholinergic system. Clearly, autistic children demonstrate a pattern of autonomic dysfunction when compared to normal children. However, the limited sample sizes employed in this preliminary experiment precluded the ultilization of the statistical techniques employed to establish discrete patterns of neurohumoral dysfunction in adult psychotics.

Reference to Figure 6 reveals another experimental comparison that may throw some light upon the ontology of adrenergic responsivity. When the course of dilatation for normal children was compared to that of normal adults, it was found that the mean pupil diameter at each interval for the children was significantly greater than that of the normal adults. Again, if it is assumed that the amount of dilatation is proportional to the amount of adrenergic mediator liberated, then it may be concluded that normal children have a significantly greater capacity for increased adrenergic outflow. An interesting difference in cholinergic reactivity was revealed when the course of constriction in normal adults was compared to that of normal children. A statistical analysis of the results presented in Figure 7 revealed that normal adults manifested significantly greater constriction than normal children. However, although the absolute level of cholinergic activity was found to be consistently greater for the normal adult, the rate of cholinergic activity was found to be significantly greater for the normal child. These results suggest that normal children and normal adults are significantly different with respect to autonomic responsivity. Furthermore, although the intermediate biochemical (genetic?) factors responsible for the change in neurohumoral activity are not as yet known, it would seem that

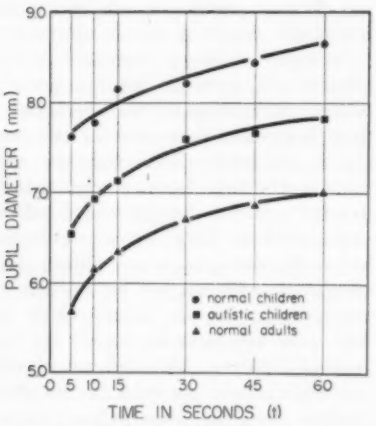


FIG. 6. Change in pupillary dilatation as a function of the amount of time spent in darkness for normal adults, normal children, and autistic, psychotic children.

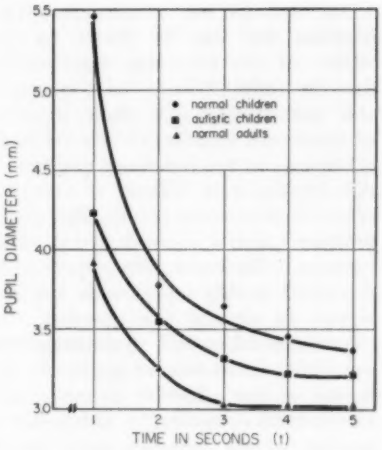


FIG. 7. Change in pupillary constriction as a function of duration of light adaptation for normal adults, normal children, and autistic, psychotic children.

the normal child, as he develops into a normal adult, undergoes changes that reduce the level of adrenergic outflow, decrease the rate of cholinergic activity, and increase the absolute level of cholinergic outflow.

SUMMARY

The results of the studies described in this report confirmed several hypotheses derived from a theoretical formulation of the role of adrenergic and cholinergic mechanisms in psychotic disorders. The salient features of the theory are: effective adjustments to emergency situations require a maximal, generalized sympathetic response; to attain a maximal level of sympathetic activity, it is not only necessary that increased levels of adrenergic outflow follow the emergency but that the level of cholinergic activity decrease; the neurohumoral dysfunctions that characterize the psychoses represent significant departures from the adrenergic-cholinergic states that characterize the normal individual at rest and during stress; and the varieties of dysfunction represent discrete levels of activity of interacting adrenergic and cholinergic mechanism. That is, each pattern of neurohumoral imbalance represents an impairment of function of either the adrenergic or cholinergic mechanism or both.

The nature of the adrenergic impairment that characterizes some psychotics may be a neurohumoral deficiency of some biogenic amine or some enzyme disorder such as monoamine oxidase overactivity in the adrenergic centers of the brain. Other psychotics manifest adrenergic impairment by a neurohumoral excess of some sympathomimetic amine that may be the result of some enzyme disorder such as monoamine oxidase underactivity in central adrenergic centers. Similarly, the cholinergic impairment of some psychotics may be characterized by excessive concentrations of brain acetylcholine or deficient acetylcholinesterase, while other psychotics may possess an inadequate level of acetylcholine or excessive acetylcholinesterase activity. A theoretical matrix of neurohumoral steady states, derived from considerations of the interaction of the levels of the two systems, characterized the population of functional psychotics as well as the normal individual at rest and under stress.

The application of measurements of pupillary reactivity to the study of autonomic dysfunction in autistic, psychotic children demonstrated that adrenergic-cholinergic imbalance clearly differentiated these children from normal children. The significance of this obtained difference was evaluated with respect to the pattern of autonomic function manifested by normal adults.

That the adult psychotic may be classified into any one of seven discrete categories of adrenergic-cholinergic activity levels throws reasonable doubt upon the current quest for a single chemical panacea. Instead, the elucidation of these discrete patterns of adrenergic-cholinergic activity suggests a rational approach to chemotherapy. Each pattern of adrenergic-cholinergic dysfunction requires a specific combination of drugs that differentially affect the levels of activity of either or both components of the autonomic nervous system.

REFERENCES

AXELROD, J. Is adrenochrome present in the blood? *Amer. J. Psychiat.*, 1958, **115**, 162–163.

BACQ, Z. M. The metabolism of adrenaline. *Pharmacol. Rev.*, 1949, **1**, 1–26.

BLISS, C. I., & CALHOUN, D. W. *Outline of biometry.* New Haven, Conn.: Yale Cooperative Corporation, 1954.

BRODIE, B. B., & SHORE, P. A. A concept for a role of serotonin and norepinephrine

as chemical mediators in the brain. *Ann. N. Y. Acad. Sci.*, 1957, **66**, 631–642.

CANNON, W. B. *The wisdom of the body.* New York: Norton, 1932.

CERLETTI, A., & ROTHLIN, E. Role of 5-hydroxytryptamine in mental diseases and its antagonism to lysergic acid derivates. *Nature, London*, 1955, **176**, 785–786.

ENGELHART, E. Der humorale Wirkungsmechanismus der Okulomotoriusreizung. *Arch. ges. Physiol.*, 1931, **227**, 220–234.

FEDEROFF, S., & HOFFER, A. Toxicity of blood serum from schizophrenics and non-schizophrenic subjects. *J. nerv. ment. Dis.*, 1956, **124**, 396–398.

FELDBERG, W., & VOGT, M. Acetylcholine synthesis in different regions of the central nervous system. *J. Physiol.*, 1948, **107**, 372–381.

FINKELMAN, I., & STEPHENS, W. M. Head regulation in dementia praecox; reactions of patients with dementia praecox to cold. *J. Neurol. Psychopath.*, 1936, **16**, 321–340.

FISCHER, R. Stress and toxicity of schizophrenic serum. *Science*, 1953, **118**, 409–411.

FORRER, G. R. Symposium on atropine toxicity therapy: History and future research. *J. nerv. ment. Dis.*, 1956, **124**, 256–259.

FROHLICH, A., & LOEWI, O. Über eine Steigerung der adrenalin Empfindlichkeit durch Cocain. *Arch. exp. Path. Pharmakol.*, 1910, **62**, 159–169.

FULTON, J. F. *A textbook of physiology.* (16th ed.) Philadelphia: Saunders, 1949.

HESS, W. R., & KOELLA, W. Experimentelle Studien über die antagonistische Innervation. *Z. ges. exp. Med.*, 1950, **116**, 431–443.

HESS, W. R., KOELLA, W., & SZABO, T. Experimentelle Studien über die antagonistische Innervation. *Z. ges. exp. Med.*, 1950, **115**, 481–490.

HOFFER, A. Epinephrine derivatives as potential schizophrenic factors. *J. clin. exp. Psychopath.*, 1957, **18**, 27–60.

HOFFER, A., OSMOND, H., & SMYTHIES, J. Schizophrenia: New approach; result of year's research. *J. ment. Sci.*, 1954, **100**, 29–54.

HOLLAND, R., COHEN, G., GOLDENBERG, M., SHA, J., & LEIFER, I. Adrenaline and noradrenaline in urine and plasma of schizophrenics. *Federat. Proc.*, 1958, **17**, 378.

KOELLE, G. B. Possible mechanisms for the termination of the physiological actions of catecholamines. *Pharmacol. Rev.*, 1959, **11**, 381, 386.

KUNTZ, A. *The autonomic nervous system.* (4th ed.) Philadelphia: Lea & Febiger, 1953.

LABROSSE, E., AXELROD, J., & KETY, S. O-methylation, the principal route of metabolism of epinephrine in man. *Science*, 1958, **128**, 593–594.

MENDEL, B., & HAWKINS, R. D. Removal of acetylcholine by cholinesterase injections and effect thereof on nerve impulse transmission. *J. Neurophysiol.*, 1943, **6**, 431–438.

OSMOND, H., & SMYTHIES, J. Schizophrenia: New approach. *J. ment. Sci.*, 1952, **98**, 309–315.

PEKKARINEN, A., GASTREN, O., & IISALO, E. Effect of iproniazid on ganglionic-blocking substances on the noradrenaline content of tissues. In, *Fourth International Congress of Biochemistry*, 1958, **15**, 106–107.

PFEIFFER, C. C., JENNEY, E. H., GALLAGHER, W., SMITH, R. P., BEVAN, W., KILLAM, K. F., KILLAM, E. E., & BLACKMORE, W. Stimulant effect of 2-dimethyl aminoethanol. *Science*, 1957, **126**, 610-611.

PFISTER, H. D. Disturbances of autonomic nervous system in schizophrenia and their relations to insulin cardiazol and sleep treatments. *Amer. J. Psychiat.*, 1938, **94**, 109–118.

Poos, F. Über die Eignung der Pupillae als Testobjekt für pharmacologische Reaktionen und Pharmakodiagnostik an Auge. *Ergbn. Physiol.*, 1939, **41**, 883–916.

RICHTER, D. (Ed.) *Schizophrenia: Somatic aspects.* New York: Macmillan, 1957.

RUBIN, L. S. The psychopharmacology of lysergic acid diethylamide (LSD-25). *Psychol. Bull.*, 1957, **54**, 479–489.

RUBIN, L. S. Acetylcholine hydrolysis in psychiatric patients. *Science*, 1958, **128**, 254–255. (a)

RUBIN, L. S. Cholinesteraese activity. *Science*, 1958, **128**, 1176–1178. (b)

RUBIN, L. S. Recent advances in the chemistry of psychotic disorders. *Psychol. Bull.*, 1959, **56**, 375–383.

RUBIN, L. S. Pupillary reactivity as a measure of adrenergic-cholinergic mechanisms in the study of psychotic behavior. *J. nerv. ment. Dis.*, 1960, **130**, 386–400.

RUBIN, L. S. Autonomic dysfunction in psychotic adults and autistic children. *Arch. gen. Psychiat.*, 1962, **7**, 1–14. (a)

RUBIN, L. S. Patterns of pupillary dilatation and constriction in psychotic adults and autistic children. *J. nerv. ment. Dis.*, 1961, **133**, 130–142. (b)

SHORE, P. A. Possible mechanism of antidepressant action of Marsilid. *J. clin. exp. Psychopath.*, 1958, **19**, 56–60.

SMYTHIES, J. R. Biochemical concepts of schizophrenia. *Lancet*, 1958, **2**, 308–313.

TRUITT, E. B. Some pharmacologic correlations of the chemotherapy of mental disease. *J. nerv. ment. Dis.*, 1958, **216**, 184–210.

VOGT, M. The concentation of sympathin in different parts of the central nervous system under normal conditions and after the administration of drugs. *J. Physiol.*, 1954, **123**, 451–481.

WOOLLEY, D. W., & SHAW, E. A biochemical and pharmacological suggestion about certain mental disorders. *Science*, 1954, **119**, 587–588.

ZELLER, E. A., BARSKY, J., & BERMAN, E. R. Amine oxidases inhibition of monoamine oxidase by 1-isonicotinyl-2-isoprophylhydrazine. *J. biol. Chem.*, 1955, **214**, 267–274.

(Received June 27, 1961)

Psychological Review
1962, Vol. 69, No. 6, 520–536

SUCCESS AND SPEED OF PROBLEM SOLVING BY INDIVIDUALS AND GROUPS [1]

FRANK RESTLE [2] AND JAMES H. DAVIS [3]

Michigan State University

The problem solving process, which begins with the presentation of a problem, is terminated when the subject arrives at a correct solution, when he arrives at an incorrect solution he believes correct, or when the experimenter arbitrarily ends the experimental session. Performance on a problem is usually summarized by the proportion of subjects who solve, or by the mean or median time to solution. A more complete description of the data gives the success of each subject (whether he obtained the correct answer or not) and the time taken; that is, a distribution of solution times or times until terminal error. One purpose of this paper is to show how the full distributions can be used to answer theoretical questions about the structure of the problem.

Small ad hoc groups of subjects can work together on verbal puzzles. Such cooperation seems natural and is appropriate to the nature of the problem and the usual behavior of college students. Generally speaking, small groups solve problems more rapidly and surely than individuals from the same population (Faust, 1959; Husband, 1940; Lorge, Fox, Davitz, & Brenner, 1958; Lorge, & Solomon, 1955; Marquart, 1955; Shaw, 1932; Taylor & Faust, 1952; Thorndike, 1938; and Watson, 1928). The advantage of groups may be a simple effect of pooling, "two heads are better than one," for in cooperative problem solving the group gets credit for the first individual insight. Other factors such as interference between group members or emergent group facilitation of thinking must be measured as deviations from the simple "pooling" hypothesis. The second aim of this paper is to formulate the pooling hypothesis so that it applies to distributions of solution times.

Model of the Distribution of Solution Times

Suppose that time is divided into short equal intervals of duration h. In any interval the subject may either solve the problem or not. Imagine that there exists a probability φ that the problem is solved in the nth interval given that it has not been solved previously. The probability φ is taken by hypothesis to be constant over intervals. Then the probability that the problem is actually solved in exactly the nth interval is the probability that it is not solved in any of the first $n - 1$ intervals (an event with probability $(1 - \varphi)^{n-1}$) times the probability that it is solved in the nth interval. That is,

$$P(n) = \varphi(1 - \varphi)^{n-1} \qquad [1]$$

the geometric distribution. This is a positively skewed (J-shaped) distribu-

[1] Developed from part of a doctoral dissertation submitted to Michigan State University by James H. Davis (1961). The first author directed the dissertation problem.

[2] Faculty Research Fellow of the Social Science Research Council. Now at Indiana University.

[3] During part of the work of this problem, Summer Fellow of the National Science Foundation. Now at Miami University (Ohio).

tion with mean $1/\varphi$ and variance $(1 - \varphi)/\varphi^2$.

To say that φ is constant over trials, that the probability of solving is independent of how many times the subject may have failed to solve the problem, is in effect to say that problem solving is a one-stage, all-or-nothing process. This assertion is frequently unreasonable in that some problems clearly seem to involve more than one stage.

Multi-stage Problems

Suppose that to obtain the final solution the subject must actually solve a sequence of stages. The stages may be solved in fixed order, or it may be that the order of stages can be inverted, but suppose that the subject actually can work on only one stage at a time. Furthermore, suppose that the time to solve an individual stage has the geometric distribution, and that the distribution of times to solve Stage 2 is independent of how long it took to solve Stage 1, etc. (Of course, it is assumed that work on Stage 2 does not even start until Stage 1 is completed.) The distribution of times to solution for a k stage problem can be determined; it is a random variable, and the sum of the k random variables which are the times to complete the k stages. If the k stages have the same probability of solution per interval, φ, then the probability that the kth stage (and the whole problem) are completed at just trial N, is given by the negative binomial (Pascal) distribution,

$$P(N; k, \varphi) = \binom{N-1}{k-1} \varphi^k (1 - \varphi)^{N-k} \quad [2]$$

An intuitive justification of this formula is given by Bush and Mosteller (1955), Ch. 14) and its derivation is discussed by Feller (1957, pp. 155–157). Note that Equation 2 reduces to Equation 1 when $k = 1$. The expected number of time intervals to solution, the expectation $\mathrm{E}(N)$, and the variance of number of intervals, are given by

$$\mathrm{E}(N) = k/\varphi \quad [3]$$

$$\mathrm{Var}(N) = k(1 - \varphi)/\varphi^2 \quad [4]$$

Although it is not likely that the various stages of a given problem would be equally difficult (have equal values of φ), there is great difficulty in estimating φ's of separate stages unless one has direct observations of the completion of various stages. Such analyses can sometimes be done but it is helpful to be able to deal directly with total solution time. For this purpose it is expedient to use the simplifying hypothesis that all stages are equally difficult, so as to be able to estimate the number of stages in the problem.

Continuous Time

The above remarks deal with short intervals of time, yet time to solve a problem may be measured continuously. If the interval h is taken smaller and smaller, then the probability of solution within interval h, $\varphi(h)$, also decreases but the ratio $\psi(h)/h$ will (in any application of the present theory) approach a limit. Let

$$\lambda = \lim_{h \to 0} \left(\frac{\varphi(h)}{h} \right)$$

and it can be shown by standard methods that for a one-stage process, the density function describing the distribution of problem solving times is

$$f(t; \lambda) = \lambda e^{-\lambda t} \quad [5]$$

the exponential distribution; and in the case of k equally difficult stages,

$$g(t; \lambda, k) = \frac{\lambda}{(k-1)!} e^{-\lambda t}(\lambda t)^{k-1} \quad [6]$$

the gamma distribution. The mean and variance of the gamma distribution are

$$E(t) = k/\lambda = \mu \quad [7]$$

$$\text{Var}(t) = k/\lambda^2 = \sigma^2 \quad [8]$$

The exponential and gamma distributions are conceptually just like the geometric and Pascal distributions given above; in fact they are limits taken by going to continuous time (intervals of duration zero).

Estimates of Parameters

A gamma distribution with parameter k (k stages) is the distribution of the sum of k exponentially distributed, independent, random variables. An exponentially distributed variable has mean equal to standard deviation, $\mu = \sigma = 1/\lambda$, from Equation 7 and 8. When one adds independent random variables, the mean of the sum is the sum of the means, and the variance of the sum is the sum of the variances. With more stages, accidental fluctuations tend to cancel out somewhat, so that the standard deviation becomes smaller relative to the mean, the more stages are involved. With increasing k, the gamma distribution becomes more nearly symmetrical, and narrower. This property can be used to estimate the parameter k. Note that from Equations 7 and 8,

$$\mu^2 = \frac{k^2}{\lambda^2}$$

and

$$\sigma^2 = \frac{k}{\lambda^2}$$

so that

$$\mu^2/\sigma^2 = k \quad [9]$$

Given large samples of observations, it is possible to substitute the obtained mean and variance, $\bar{t}$ and s_t^2 for the parameters, obtaining the estimation equation

$$\hat{k} = (\bar{t})^2/s_t^2 \quad [10]$$

The authors have used this estimate, an example of the method of moments, instead of possible maximum likelihood or minimum chi square estimates. The method-of-moments estimate seems relatively insensitive to imperfections in the experimental procedure, and its general significance remains sensible even if the actual stages of the problem are not exactly equal in difficulty. If the stages differ in difficulty, then $\hat{k}$ can be considered as follows; the actual problem is represented by a model problem with equally difficult and independent stages, and k is the number of stages in the model. The interpretation is like one common in information theory; if a certain system transmits just 3 bits, it can be represented by a system which perfectly discriminates exactly 8 equally probable signals. The value of 3 bits may actually be derived from a system which makes imperfect discriminations of 24 categories which are not equally probable, but the model gives a general impression of the behavior of the system. Similarly, our use of a gamma distribution gives a general impression of the number of stages in a problem, of its "complexity," without necessarily being correct about the detailed structure of the problem.

Application

Davis (1961) gave three problems to 178 undergraduates, and timed each subject on each problem. The order of the three problems was changed from subject to subject. A

sizable fraction of the subjects arrived at wrong answers on some problems, and the wrong answers will be set aside; even if there are exactly k stages to solution of the problem, there is no assurance that it also takes k stages to arrive at a wrong answer.

On the basis of subsidiary measurements, a relatively constant "reading time" could be subtracted from the total time consumed on each problem. When this was done, means and standard deviations of net solution time were calculated on each of the three problems.

The first (Rope) problem described a prisoner attempting to escape from a tower, who found in his cell a rope which was half long enough to permit him to reach the ground safely. He divided the rope in half and tied the two parts together and escaped. How could this have been? The answer is that the rope is divided lengthwise. This problem was taken to be a "one-stage" problem in that the solution seems to depend on a single idea.

The net mean time was 131.3 seconds, the standard deviation was 115.5, whence the estimated number of stages is $(131.3)^2/(115.5)^2 = 1.3$ stages, not far from the one stage predicted. In fact, it is possible that the time taken to write the solution might produce the modest discrepancy.

The second (Word Tangle) problem was a more complex one: "If the puzzle you solved before you solved this one was harder than the puzzle you solved after you solved the puzzle you solved before you solved this one, was the puzzle you solved before you solved this one harder than this one?" Subjects were to answer yes or no and why. The answer (yes) requires canceling of several mutually defeating phrases. Net mean time was 264.8 seconds, with a standard deviation of 154.1, so that the estimate of k is 3.0 stages.

The third (Gold Dust) problem was a more complex version of the standard water-jug problem. Given containers of 163, 14, 25, and 11 ounces, obtain exactly 77. Solutions are to subtract, from 163, either $2(25) + 14 + 2(11)$, or $25 + 2(14) + 3(11)$, i.e., either five or six manipulations. The net mean time was 373.0 seconds, standard deviation 166.7, whence the estimated number of stages is $\hat{k} = 5.0$.

The estimates of number of stages of of these problems seem, to the present writers, quite in accord with our intuitive judgments of number of stages. When a group of students, after having solved the problems, were asked to estimate the number of stages in the three problems, their mean judgments were 2.8, 4.0, and 6.3, respectively. These judgments are consistently higher than those obtained from the distributions of solution times, but the correlation is remarkable and indicates that our judgments as to the relative number of stages in the three problems agree in substance with judgments of student solvers.

In addition, it was found that the gamma distributions (Equation 6) give satisfactory fits to the obtained distributions of solution times on the three problems. The obtained cumulative distributions of time to correct solution of the three problems are shown in Figure 1, along with fitted gamma distributions. An investigation of overall goodness of fit of the theoretical gamma distribution was made separately for each problem, using the Kolgomorov-Smirnov one-sample test. The maximum discrepancies between theoretical and observed distributions were all well within the .05 acceptance region for all three problems.

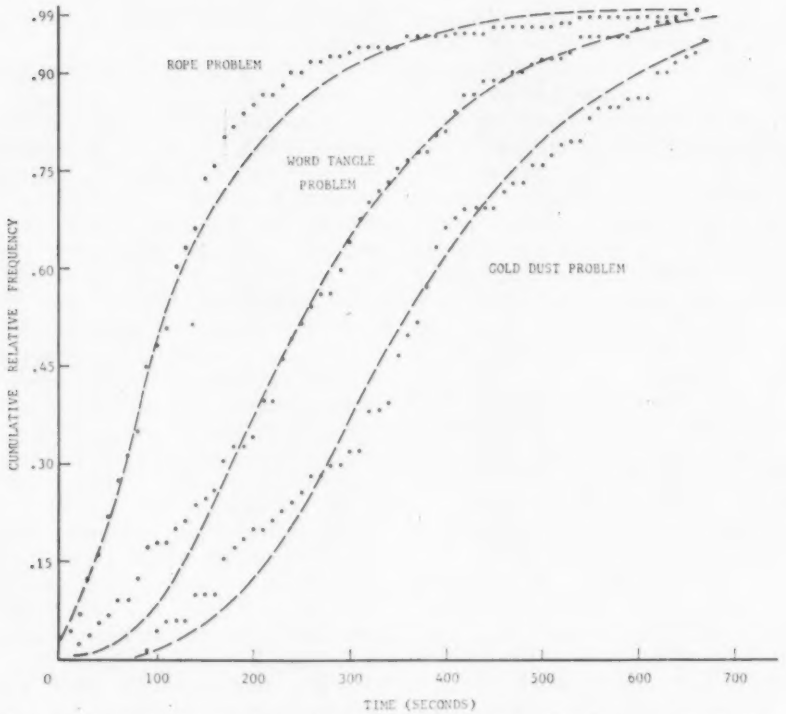


FIG. 1. Cumulative relative frequency of individual solvers and fitted gamma distributions with k = 1.3, 3.0, and 5.0 for Rope, Word Tangle, and Gold Dust Problems, respectively.

Meaning of the Model for Distribution of Solution Times

The model for the distribution of solution times assumes that the times subjects take to solve a problem vary widely because solving a problem involves having only one or a few ideas, in sequence; and the time to an idea is taken to be a random variable. This seems a natural consequence of thinking of such a problem solving process as "insightful." The detailed model includes other, less attractive, simplifying assumptions. One is that the stages of a problem are all equally difficult. Another is that the same parameter λ applies to all members of the population tested. These assumptions are probably wrong in detail, though small deviations may not be easily detectable from the kind of data we have used. It is emphasized that they are only simplifying assumptions which are made, in this model, because we were not able to separate and measure the stages of solution of the problems (because we did not, originally, know what they might be, and we are not sure now), and we did not have good individual differences predictors for our subjects.

POOLING OF ACCOMPLISHMENTS IN GROUPS

In the usual experiment on problem solving by groups, groups are better

than individuals. If "Eureka" problems are used, the group is likely to solve successfully as soon as the first group member arrives at the correct solution. If it is assumed that group members are just as effective in groups as they would be working alone, then it is possible to predict group from individual performance, provided that individuals and members of group are sampled randomly from the same population of subjects, and groups are formed randomly.

The one model available for such predictions is by Lorge and Solomon (1955) who divide the population into "solvers" and "nonsolvers," use tests of individuals to estimate P_I (the proportion of solvers among individuals) and then predict that the probability of finding at least one solver in a group of size r will be $P_r = 1 - (1 - P_I)^r$. Using this theory it is possible to transform the cumulative distribution of individual solvers, which gives P_I as a function of t, into a cumulative distribution for groups of size r, though Lorge and Solomon do not do so.

Lorge and Solomon also note that if a problem has more than one part or stage, it is possible that a group will solve it though no individual in the group can solve it. Different group

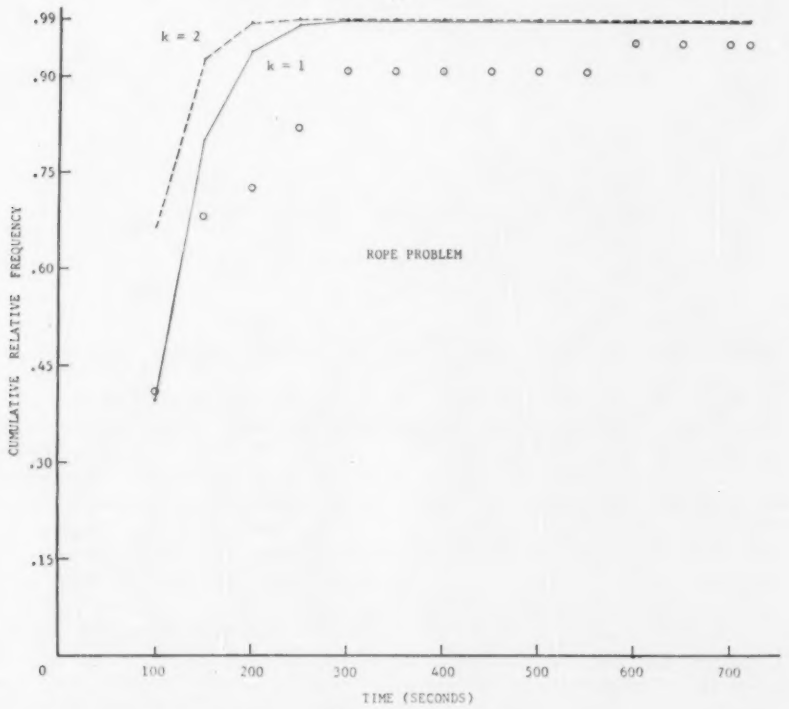


FIG. 2, Part 1.

FIG. 2. The solid and broken lines represent cumulative relative frequency at 50-second intervals, predicted by Lorge-Solomon Model ($k = 1$, $k = 2$, respectively) for groups of Size 4. (Observed group performance, for the same intervals, is given by solid points.)

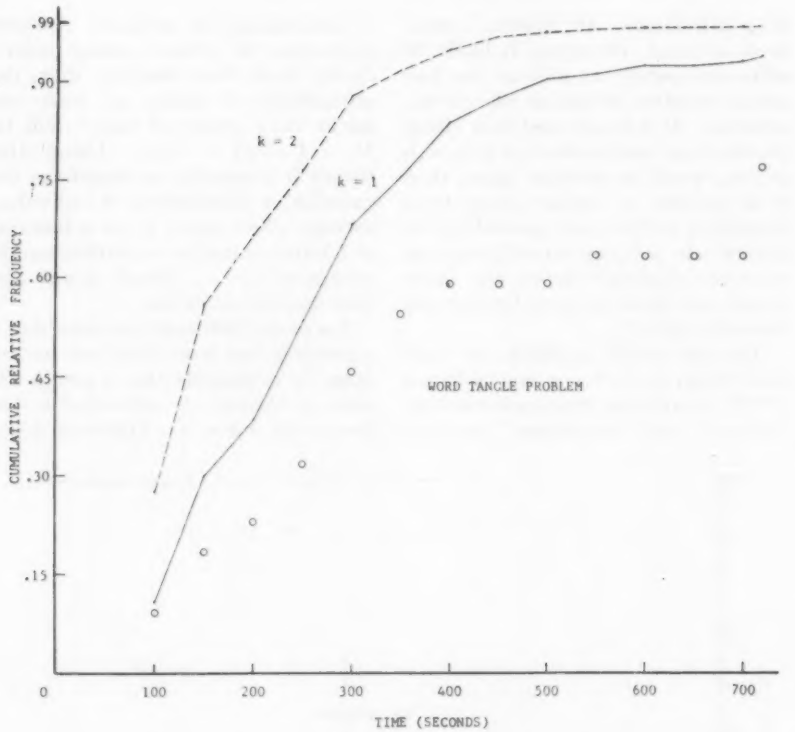


FIG. 2, Part 2.

members may be able to solve different parts, and then pool their abilities to obtain an emergent solution. If the abilities to solve the various parts are equally frequent in the population and are independent, this gives a decided extra advantage to groups. With k equally probable stages the probability of a group solution with r people is:

$$P_r = [1 - (1 - P_1^{1/k})^r]^k \quad [11]$$

Equation 11 also can be used to generate a distribution of group solution times given a distribution of individual solving times, for any given value of k.

The use of the waiting-time model for the distribution of solution-times yields an estimate of k for a problem, from the distribution of individual solution times. That estimate can be inserted for k in Equation 11, and group performance can be predicted from individual performance. Davis (1961) performed the appropriate experiment. In addition to the 178 individuals, he had 22 four-person groups work on his three problems. Recruitment of subjects, working conditions, instructions, etc., were closely comparable for individuals and groups. When Equation 11 is used to predict group from individual solution times, the results are very inaccurate. Figure 2 shows the predicted group

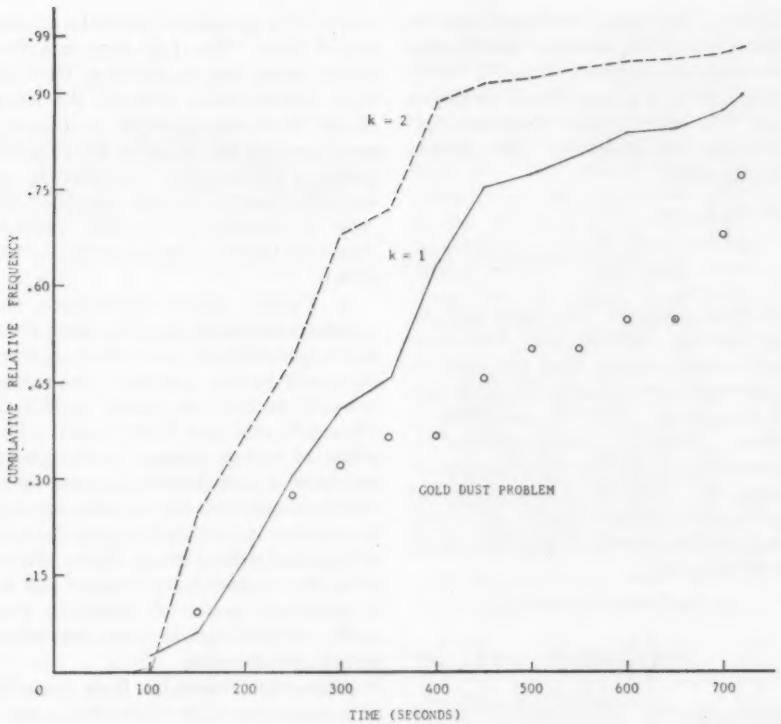


FIG. 2, Part 3.

performance (from Equation 11) and the obtained distributions on the Rope, Word Tangle, and Gold Dust problems for $k = 1$ and $k = 2$. The actual estimates of k were 1.3, 3.0, and 5.0, which would obviously give extremely inaccurate estimates.

However, deeper consideration shows that the Lorge-Solomon model is not the proper one to be used with the waiting-time model developed. Lorge and Solomon begin by dividing the population into solvers and nonsolvers. To obtain more complete predictions, we extended the idea to "solvers by time t" and "nonsolvers at time t," which seems a legitimate extension of the ideas of Lorge and Solomon. However, we earlier had generated the distributions of solution time from the assumption that solution was a random event, and all subjects had equal probability of solution in a given interval. The Lorge-Solomon model takes the individual performance as a fixed fact which identifies solvers and nonsolvers, assumed to be individuals with different abilities. The random waiting-time model leads to a different, combination-of-*contributions* model.

Combination of Contributions

Consider a group of subjects none of whom misinterprets the problem or arrives at a wrong solution. Suppose that for each subject the problem

involves the same k stages each of which has an exponential distribution of time-to-completion of $\lambda e^{-\lambda t}$. With r subjects in a group, it can be shown that the distribution of group performance is given by the gamma distribution,

$$g(t; r\lambda, k) = \frac{r\lambda}{(k-1)!} e^{-r\lambda t}(r\lambda t)^{k-1} \quad [12]$$

which is precisely the same gamma distribution derived for individual performance except that the rate (λ) is increased by a factor of r, the size of the group. Part of the proof is given. Within a short interval h, given that no one has solved a certain stage, the probability that some one in the group will solve is $1 - (1 - \Psi)^r$. As h is taken small, Ψ becomes small, in which case

$$1 - (1 - \Psi)^r = 1 - 1 + r\Psi - \binom{r}{2}\Psi^2 + \cdots \pm \binom{r}{r}\Psi^r$$

approaches $r\Psi$. The reason is that all terms involving Ψ^2, Ψ^3, etc., become relatively negligible when Ψ becomes small. Hence, in the limit, the distribution of time to a given stage approaches $r\lambda$.

This calculation provides a prediction of group problem solving if no subjects arrive at wrong solutions. Preliminary results of an experiment by R. Hoppe at Michigan State University[4] give some support for this model. Hoppe used Rokeach's (1960) "Doodlebug" problem and tested individuals, and groups of Size 2 and 3. In Hoppe's study the subjects could ask the experimenter questions and were, in the main, redirected when they had made a serious mistake, so that all subjects and groups either solved the problem correctly or ran out of time. The data were in agreement with the hypothesis that all time distributions were of the same shape (it is not possible to make a good test of the gamma distribution since a substantial proportion of subjects failed to finish) and that the rate of solution is roughly proportional to the number of people in the group.

In Davis' (1961) experiment the subjects received no redirection from the experimenter, and about half of the individuals arrived at wrong answers on the two harder problems (Word Tangle and Gold Dust). The effect of wrong answers in the group situation is not directly known, since observations did not enable the experimenter to tell which members of groups had solved the problem. However, two plausible hypotheses can be formulated, and they lead to two quite distinct predictions regarding group performance.

Hierarchical model. It is possible that subjects going "off the track," referred to here as nonsolvers, are nonfunctional in their groups. If so, a group with two solvers and two nonsolvers will solve the problem just as rapidly, on the average, as a group of two solvers alone. (It is assumed that in Eureka problems, groups will not agree on a wrong solution.) Nonsolvers would be nonfunctional if, for example, the solvers suppress nonsolvers and form an intellectual hierarchy within the group.

If a group of size r has A solvers and $B = r - A$ subjects who are nonsolvers, then by the Hierarchical model the distribution of time to solution should be a gamma distribution, $g(t; A\lambda, k)$. If a proportion a of individual workers solve the problem, then the probability that a group of size r has exactly A members

[4] Ronald Hoppe, personal communication, 1961.

who will solve is

$$P(A) = \binom{r}{A} a^A (1 - a)^{r-A}, \quad [13]$$

a binomial distribution.

The distribution of solution times of a collection of groups will then be described by the density function,

$$f_H(t) = \sum_{A=0}^{r} P(A) g(t; A\lambda, k) \quad [14]$$

if the hierarchical model is correct.

Equalitarian model. Another possibility is that those subjects who go off the track, the nonsolvers, continue to take full part in group discussion though they do not arrive at solutions. This idea leads to the hypothesis that the nonsolvers would consume their share of the group's time, slowing up the other members of the group. This would mean that the probability that a solver in the group solves in unit time is reduced from λ to $[A/(A + B)]\lambda$, when there are A solvers and B nonsolvers present. The distribution of solution times for such a group is, by the argument leading to Equation 12,

$$g\left(t; A\,\frac{A}{A+B}\lambda, k\right) = g\left(t; \frac{A^2}{A+B}\lambda, k\right) \quad [15]$$

Given that the probability of an individual's solving is a, the distribution of solution times of a collection of groups will be described by the density function,

$$f_E(t) = \sum_{A=0}^{r} \left[P(A) g\left(t; \frac{A^2}{A+B}\lambda, k\right)\right] \quad [16]$$

if the equalitarian model is correct.

The evidence of Davis' (1961) experiment is that the hierarchical model is seriously in error, and the equalitarian model yields predictions which are quite close to the data, not seriously in error. The Kolgomorov-Smirnov one-sample test leads to rejection of the distribution calculated from the Hierarchical Model, for the word tangle and gold dust problems, at the .05 level of significance. The distribution derived from the Equalitarian Model is accepted at the .05 level for all three problems. Of course, with only 22 groups, the statistical acceptance is not decisive. Predictions from these two combination-of-contributions models, along with the distributions of group performance, are shown in Figure 3.

Subsidiary evidence in the form of sociometric choices and communication frequencies (Davis, 1961) was collected, and the results agree in substance with the assumption of an equalitarian group structure.

DISCUSSION

Structure of Problems

The estimates of number of stages of the Rope, Word Tangle, and Gold Dust problems would be very impressive evidence, except that we have no strong independent measurement of the number of stages in such problems. The intentions and judgments are suggestive but not decisive. The number of stages in a problem is the number of distinct ideas required to obtain the solution. The number of stages is the number of ideas which must be conjoined to obtain the solution—the subject must have Idea 1 *and* Idea 2 *and* . . . , etc. If there were alternative answers, the subject could have Idea 1 *or* Idea 2 *or* . . . , and these would not be stages. This analysis makes it possible to compare stage-estimates from the data with the logical structure of the problems; and

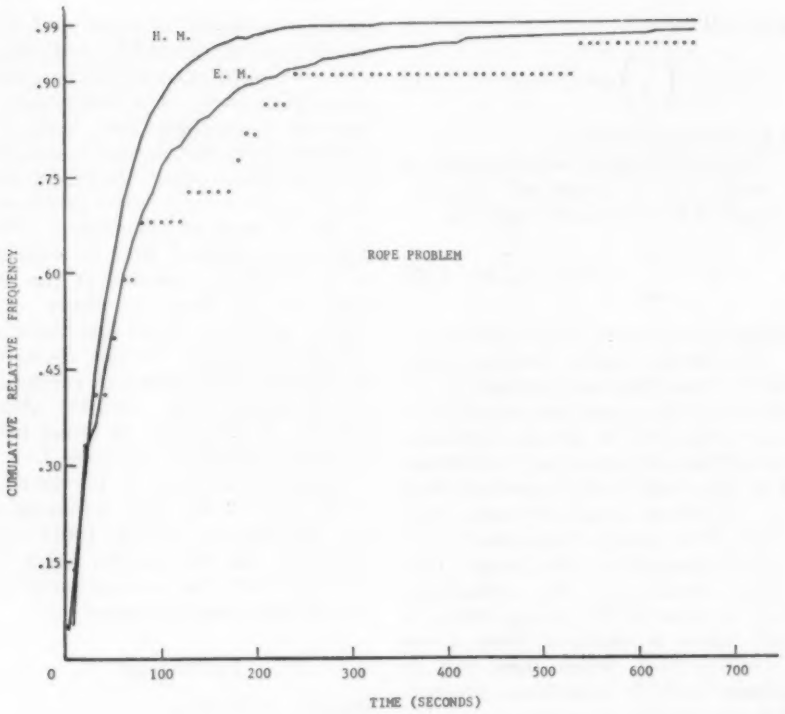


FIG. 3, Part 1.

FIG. 3. The solid lines represent cumulative relative frequency predicted by the Hierarchial and Equalitarian Models for groups of Size 4. (Observed group performance is given by solid points.)

the agreement is good. Unfortunately, we are not sure how to identify "ideas" in the sense used above, hence the logical analysis is not sure of its elements.

What is needed, obviously, is an experimental separation of stages in problem solving. Hoppe's work, mentioned above, attempts such an analysis. A difficulty is that one can hardly separate the stages of a problem, experimentally, unless he knows what those stages are. Verbal protocols of problem solving process are useful for identifying parts, but they present difficulties of interpretation. The writers do not foresee any easy and immediate answer to the question of problem structure.

Stages versus parts. In the special model above it was assumed that the stages are to be accomplished in sequence. If there are simply k parts which must be accomplished, in any order, then the distribution of times may be changed. If the subject is thought to work on only one stage at a time, solving the stages in some irregular order, then the gamma distribution model applies. However, if ideas for different parts of the problem can arise freely and in-

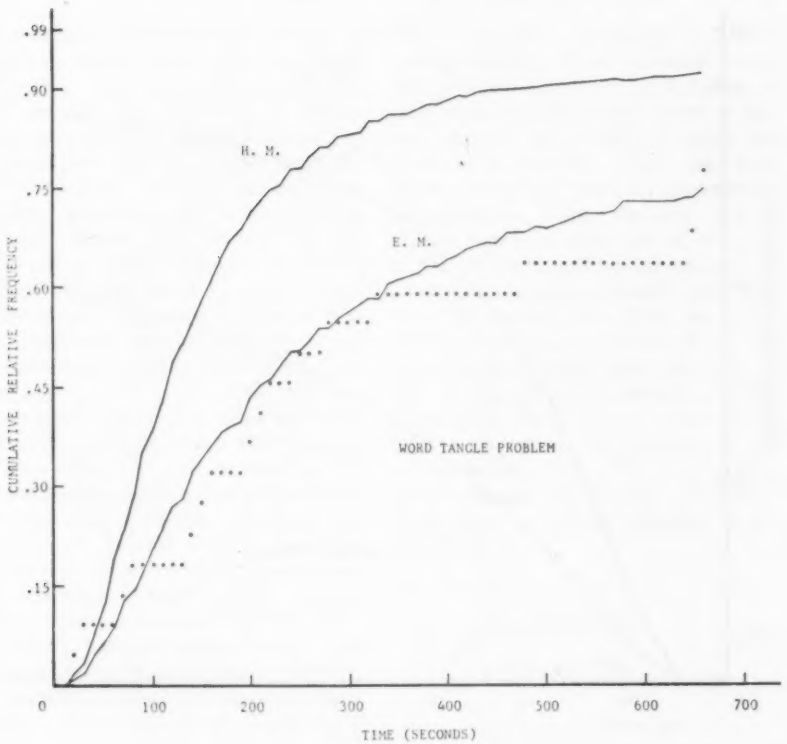


FIG. 3, Part 2.

dependently, a new distribution arises. Solution time of any part will be exponentially distributed, but until the subject has solved the first part there are k parts which may be solved. Hence the rate until the first part-solution is $k\lambda$. After one part is completed, there are $k-1$ parts left and the rate until the second part is $(k-1)\lambda$. Finally, when there is only one part left, the rate is λ. These "rates" can be defined as above if the distribution of times until the next stage is assumed to be exponential.[5]

The resulting distribution is rather complicated, but its mean and variance can be developed easily. Since the mean of each of the waiting times is independent of the others by hypothesis, the total mean is the sum of k waiting times with means $1/\lambda$, $1/(2\lambda), \cdots, 1/(k\lambda)$. Hence the mean total time to solution, $E(T)$, is

$$E(T) = (1/\lambda) \times (1+1/2+1/3+\cdots+1/k) \quad [17]$$

By a similar argument, the variance of total times is

$$\mathrm{Var}(T) = (1/\lambda^2) \times [1+(1/2)^2+(1/3)^2+\cdots+(1/k)^2] \quad [18]$$

[5] The writers are indebted to Herman Rubin for the formulation and solution of this mathematical problem.

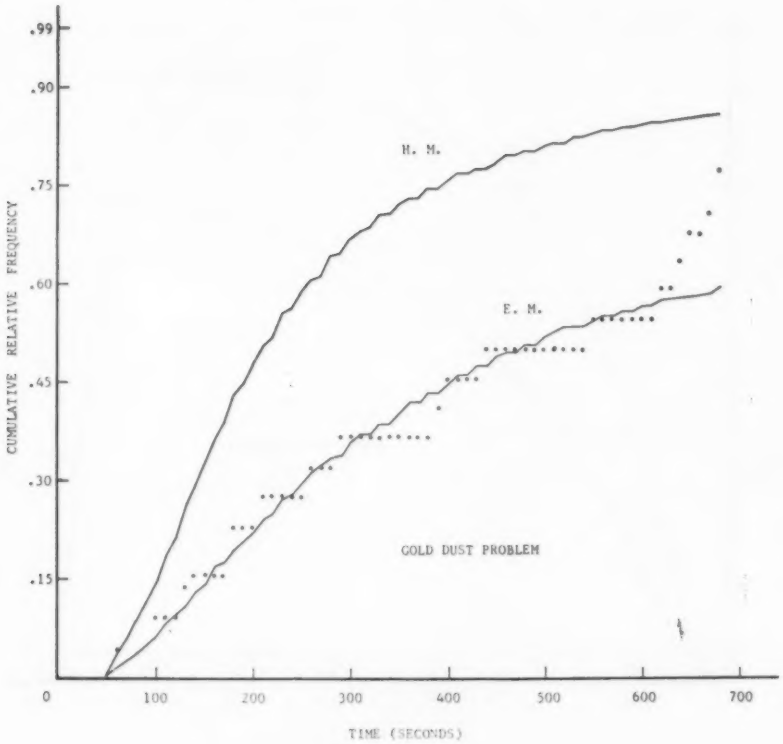


FIG. 3, Part 3.

TABLE 1

SOME STATISTICS OF PROCESSES WITH PARTS (Which May be Solved Concurrently) AND PROCESSES WITH STAGES (Which are Solved Successively)

Number of parts or stages	$\mu\lambda$		$\sigma\lambda$		μ^2/σ^2	
	Parts	Stages	Parts	Stages	Parts	Stages
1	1.00	1.00	1.000	1.000	1.000	1.00
2	1.50	2.00	1.108	1.414	1.800	2.00
3	1.83	3.00	1.117	1.731	2.469	3.00
4	2.08	4.00	1.119	2.000	3.049	4.00
5	2.28	5.00	1.210	2.238	3.562	5.00
6	2.45	6.00	1.221	2.452	4.025	6.00
7	2.59	7.00	1.229	2.642	4.447	7.00
8	2.72	8.00	1.236	2.828	4.836	8.00
9	2.83	9.00	1.241	3.000	5.198	9.00
10	2.93	10.00	1.245	3.161	5.536	10.00

Table 1 shows how mean and standard deviation, and also the value of μ^2/σ^2, progress with more and more parts. Comparable values for "stages," i.e., parts which must be solved in succession, are shown for contrast.

Possible effects of wrong solutions. The Equalitarian model for predicting group from individual performance was chosen because it seemed to take proper account of the fact that some subjects may arrive at wrong but self-satisfying answers. However, wrong solutions may distort the distribution of times to correct solution. In estimating the number of stages to solution and working out the waiting-

time model, subjects who arrived at wrong solutions were eliminated from the data. This seems reasonable, for the number of stages to a wrong solution may be different from the number of stages to a correct solution. However, this conditional distribution of times (times to solution given that a correct solution is reached) may be misleading. The difficulty is that a subject who happens to take a long time to get the correct answer has an increased vulnerability to error, whence long times to correct solution may be systematically less frequent when errors occur than they would be if the whole distribution were known.

The difficulties can be studied by investigating the properties of random systems with more than one outcome. The most general model available to the present writers is that proposed by LaBerge (1960)[6] which permits any of several responses which involve k_1, k_2 . . . stages respectively. LaBerge derives the distribution of times to completion of the k_1 stages of Outcome 1 provided that Outcome 1 occurs.

One special model may be called the "misreading" model, in which it is supposed that the subject may, at the very beginning, make a choice which leads to a wrong solution. The subject either enters into a process which leads to success (with the gamma distribution of Equation 8), or onto some other which leads to failure. The conditional distribution of times to correct solution, given that a correct solution and not an error is attained, is the original gamma distribution.

A third possibility is that the subject may at any time before solution make a decision which leads to a wrong answer. Consider that this wrong decision is a unitary event which has a constant probability of happening at any time, given it has not already happened, up until the problem is solved. Then the accidental event will have an exponential distribution. The density (probability per time interval) of solution of the problem will be the product of the original gamma density, $g(t; \lambda, k)$ and the probability that the accidental event has not occurred. If the constant probability-per-unit-time or rate of the accidental event is α, the probability that it has not occurred by time t (provided the problem is not solved before t) is $e^{-\alpha t}$. The conditional density of solution at time t, given that a correct solution is attained, is

$$f(t; \lambda, \alpha, k) = \frac{g(t; \lambda, k)e^{-\alpha t}}{\int_0^\infty g(t; \lambda, k)e^{-\alpha t}dt} \quad [19]$$

$$= \frac{\frac{\lambda}{(k-1)!}e^{-(\lambda+\alpha)t}(\lambda t)^{k-1}}{\int_0^\infty \frac{\lambda}{(k-1!)}e^{-(\lambda+\alpha)t}(\lambda t)^{k-1}dt} \quad [20]$$

The numerator is simply

$$g(t; \lambda+\alpha, k)\left(\frac{\lambda}{\lambda+\alpha}\right)^k \quad [21]$$

and the denominator is

$$\left(\frac{\lambda}{\lambda+\alpha}\right)^k \int_0^\infty g(t; \lambda+\alpha, k)dt = \left(\frac{\lambda}{\lambda+\alpha}\right)^k \quad [22]$$

The step from Equation 20 to Equation 21 is accomplished by noting that $g(t; \lambda + \alpha, k)$

$$= \frac{\lambda+\alpha}{(k-1)!}e^{-(\lambda+\alpha)t}[(\lambda+\alpha)t]^{k-1}$$

$$= (\lambda+\alpha)^k e^{-(\lambda+\alpha)t}t^{k-1}$$

[6] LaBerge, D. A statistical theory of discrimination behavior. (Prepublication draft, December 1960.)

If this is multiplied by $[\lambda/(\lambda + \alpha)]^k$ one obtains the numerator of Equation 20.

Hence the conditional distribution of solution times is

$$f(t; \lambda, \alpha, k) = g(t; \lambda + \alpha, k) \quad [23]$$

a gamma distribution with the original number of stages, k, but with rate increased to $\lambda + \alpha$. The probability of a correct solution is $[\lambda/(\lambda + \alpha)]^k$.

In summary, three effects of wrong solutions were mentioned; they might distort the distribution from a gamma to some other distribution (if the subject works concurrently on stages of several different processes leading to several answers), it might leave the conditional distribution of solution times unchanged (if errors arise only through an initial wrong decision such as misreading the problem), or it might leave the conditional distribution a gamma, still with k stages, but increase the rate of solutions, (if entry onto an erroneous path is thought of as a unitary random event.) Of these, the third seems most in accord with the problem solving situation in Davis' (1961) experiment.

Unhappily, we do not have the data needed to decide which of these possibilities, if any, occurred in the Davis (1961) experiment. As we understand it, the LaBerge formulation could be fit to the data exactly only if we knew what all the wrong solutions were and could estimate the number of stages and the rate parameters of each. Davis' experiment was not designed to facilitate such an investigation, and the necessary analysis is not feasible. Since LaBerge's model involves parameters we cannot estimate, it cannot be compared with the other models validly. The misreading model and the accidental event model do not require analysis of the inner structure of erroneous solutions, hence can be applied to the present data. Since both lead to gamma distributions of correct solution, both agree with Davis' results. Only a detailed qualitative analysis of protocols, or a technique for eliminating wrong solutions without otherwise disturbing performance, would indicate which model is correct. In fact, some subjects might misread and others err through an accidental random event, and the conditional distribution of correct solution time would still be a gamma distribution.

This theoretical result increases the plausibility of the gamma distribution as an hypothesis by showing that it might, under reasonable conditions, characterize data from problems with quite different frequencies of wrong solution.

OTHER MODELS FOR OTHER KINDS OF GROUPS

The predictions of group performance in Davis' experiment were for ad hoc groups of equivalent members, with no discernable social inequalities. Furthermore, the nature of the problems was such that ideas were pooled with little controversy over the correctness of a correct proposal. In a "team" it sometimes happens that all members must solve their parts of the problem before the group succeeds. Then the probability of group solution is the joint probability that all members have completed their tasks by time t. If the individual assignments are unitary problems with the same difficulty, and all members of the group can work concurrently (as in servicing an automobile) then the model leading to Equation 17 and 18 should describe the situation. If one stage of the process must wait upon completion of another stage, then combining the individual contributions can lead to a gamma distribu-

tion. If solution of the problem is mainly a matter of decision rather than invention, then the probability of decision by time t may be the probability that all group members accept the same idea at that time. Changes of opinion, persuasion, and the invention of compromises are the events of primary interest, and the main results of the present paper do not apply.

These remarks emphasize that the mathematical theory used in this paper is not a general theory of group problem solving, but is a careful and precise formulation of some hypotheses about one particular kind of group problem solving. The hypotheses depend both upon the constitution of the groups and the nature of the problems.

RELATION TO OTHER AREAS OF PSYCHOLOGY

The mathematical forms discussed in this paper have applications elsewhere in psychology. The notion that an individual stage of problem solving will have an exponential distribution of solution time, is essentially the same as the hypothesis that a unitary learning event will occur on a geometrically distributed trial. Both arise from the statement that the event has constant probability of occurring, until it does occur. In learning theory, this is the hypothesis of "all-or-nothing" learning (Bower, 1962; Estes, 1960; Restle, 1961b, 1962). A truly unitary process, indivisible and independent of other processes, should have a constant probability of occurrence. If an event is not unitary but a compound, then every effort should be made to tease apart the stages involved in it. Hence the geometric and exponential distributions (for trials and continuous time, respectively) give a theory for any event which can properly be left unanalyzed.

The only compounds which do not require detailed analysis into parts are those made up of equivalent parts. A beam of light can be thought of as a compound of unitary emissions of photons, but the only interesting question about such a beam is the number of photons, since separate photons are all alike. Thus, an ideal compound event must be one made up of equivalent, identically distributed parts. The gamma distribution describes compounds of the type used in this paper. Previous applications of gamma distributions include Bush and Mosteller (1955, Ch. 14) on the time rats take to traverse a runway, and Restle (1961a) on simple reaction time.

Similarly, the simplest kind of model for comparing group with individual performance is one which assumes that people in a group are, in all ways possible, just like people working alone. In a nonevaluative sense one can say that the Lorge-Solomon models, and the variations given in the present paper, are all versions of this ideal hypothesis.

Any deviations between model and data suggest the action of factors more complex (and perhaps more interesting) than the very simplest possibilities discussed here. For example, since Davis' groups did not reach the level predicted by Lorge-Solomon models, it became possible to introduce the hypothesis that individuals who have entered on a wrong interpretation of the problem take up their share of group time and interfere with the progress of others. Discrepancy from one model became a measurement of a previously unmeasured factor.

SUMMARY

A waiting-time model was developed to give a distribution of times to solution, for word puzzles. The distribution depends upon the number of stages involved in solving the problem. Experimental results showed that the theoretical distribution (gamma distribution) gives a good fit to the data, and the estimated number of stages agrees well with intuition and judgments. A theory of "pooling of contributions" was developed to describe the superiority of groups over individuals in solving word puzzles. Experimental results suggested that subjects who have misunderstood the problem and will arrive at wrong answers, nevertheless consume their share of group time. With this addition, the pooling of contributions theory accurately predicts group problem solving data. Some alternative models, resulting from different problem structures and different kinds of groups, were discussed, and the waiting-time theory was shown to be related to the concept of "all-or-nothing" learning.

REFERENCES

BOWER, G. H. An association model for response and training variables in paired associate learning. *Psychol. Rev.*, 1962, **69**, 34–53.

BUSH, R. R., & MOSTELLER, F. *Stochastic models for learning.* New York: Wiley, 1955.

DAVIS, J. H. Models for the classification of problems and the prediction of group problem-solving from individual results. Unpublished doctoral dissertation, Michigan State University, 1961.

ESTES, W. K. Learning theory and the new "mental chemistry." *Psychol. Rev.*, 1960, **67**, 207–223.

FAUST, W. L. Group versus individual problem solving. *J. abnorm. soc. Psychol.*, 1959, **59**, 68–72.

FELLER, W. *Introduction to probability theory and its applications.* New York: Wiley, 1957.

HUSBAND, R. W. Cooperative versus solitary problem solution. *J. soc. Psychol.*, 1940, **11**, 405–409.

LORGE, I., FOX, D., DAVITZ, J., & BRENNER, M. A survey of studies contrasting the quality of group performance and individual performance, 1920–1957. *Psychol. Bull.*, 1958, **55**, 337–372.

LORGE, I., & SOLOMON, H. Two models of group behavior in the solution of eureka-type problems. *Psychometrika*, 1955, **20**, 139–148.

MARQUART, DOROTHY I. Group problem solving. *J. soc. Psychol.*, 1955, **41**, 103–113.

RESTLE, F. *Psychology of judgment and choice.* New York: Wiley, 1961. (a)

RESTLE, F. Statistical methods for a theory of cue learning. *Psychometrika*, 1961, **26**, 291–306. (b)

RESTLE, F. The selection of strategies in cue learning. *Psychol. Rev.*, 1962, **69**, 329–343.

ROKEACH, M. *The open and closed mind.* New York: Basic Books, 1960.

SHAW, MARJORIE E. Comparison of individuals and small groups in the rational solution of complex problems. *Amer. J. Psychol.*, 1932, **44**, 491–504.

TAYLOR, D. W., & FAUST, W. L. Twenty questions: Efficiency in problem solving as a function of size of group. *J. exp. Psychol.*, 1952, **44**, 360–368.

THORNDIKE, R. L. On what type of task will a group do well? *J. abnorm. soc. Psychol.*, 1938, **333**, 408–412.

WATSON, G. B. Do groups think more efficiently than individuals? *J. abnorm. soc. Psychol.*, 1928, **23**, 328–336.

(Received July 28, 1961)

Psychological Review
1962, Vol. 69, No. 6, 537–541

THEORETICAL NOTES

THE MOLECULAR BASIS OF MEMORY AND LEARNING

MICHAEL H. BRIGGS AND G. BARRIE KITTO

Victoria University of Wellington, New Zealand

In recent papers, Gaito (1961) and Hyden (1961) have suggested that nerve activity results in a rearrangement in the sequence of monomers of certain high molecular weight polymers (nucleic acids and proteins) found in brain cells, or in the synthesis of additional polymers of novel sequences, and that this rearrangement or synthesis is the molecular basis of memory. It is the purpose of this note to show that the only mechanisms whereby such rearrangements or syntheses could occur predict chemical brain changes incompatible with present experimental evidence, and, moreover, that current biochemical knowledge of nucleic acids renders the hypothesis almost certainly false. The evidence brought forward in support of polymer sequence rearrangement is then shown to be compatible with a second hypothesis that is not contrary to current biochemical views on biopolymers.

There are three main types of polymers [1] in all cells, including neurones. The deoxyribonucleic acids (DNAs) are confined to the nucleus and, in the form of nucleoproteins, are the major constituents of the genes. There are two types of ribonucleic acids (RNAs). A soluble form (messenger-RNA) is apparently synthesized in the nucleus and then transferred to the cytoplasmic ribosomes. These structures contain insoluble RNAs and are the site of protein synthesis.

The interrelations of DNA, RNA, and protein are relatively simple. Every cell contains at least one type of DNA molecule for every gene. These molecules are chemically identical to the DNAs of the parent cells. Each DNA molecule is a long, unbranched chain of nucleotides, which usually occurs chemically bonded to a second molecule in the form of a twin helix. Replication of a DNA molecule occurs only at cell division when the two components of a helix separate and each serves as a template for the synthesis of a new polynucleotide to form two new helices.

As these processes occur only at cell division, it is clear that the DNA molecules in the nucleus of a neurone are permanent structures, and there is no experimental evidence of any significant change in quantity with time.[2] The base sequence of any particular DNA molecule is determined by the base sequence of the template on which it was formed (Kornberg, 1960). It has been demonstrated that no polynucleotide synthesis occurs in the absence of a template (Kornberg, 1960).

This evidence rules out changes in the base sequence of DNAs as molecular "memory banks." There is no way in which electrical changes in the cell walls of neurones could interfere with the process of polynucleotide synthesis to give new base sequences. DNA is formed only on a template, spontaneous coupling of nucleotides does not occur. Hence, even if DNA molecules of the nucleus were fragmented by an electrical impulse, they would not recombine into polynucleotides of different base sequence.

The only known mechanism for altering the base sequence of a DNA molecule

[1] Details of the biochemistry of DNA, RNA and their relation to protein biosynthesis can be found in numerous publications, e.g. Abrams (1961), Leslie (1961).

[2] Total brain DNA reaches a minimum in man at about age 30 years, then steadily increases. Changes in individual brain cell DNA concentrations are at present uninvestigated.

is by reaction with a mutagenic agent. These are compounds which chemically alter the purines and pyrimidines of the polynucleotide (see, for example, Freese, 1961). The effect of a mutagen is to alter the base sequence *not* by a rearrangement of the existing bases, but by the conversion of some of the bases *in situ* on the polymer molecule to other related compounds, not normally found in DNAs. It could be hypothesised that nerve impulses produce chemical mutagens *in vivo*. However, this cannot be true for no unusual bases have been found in brain DNAs, nor incidentally, in brain RNAs.

Nuclear DNAs continually serve as templates for the synthesis of messenger-RNA. The base sequence of the DNA template determines the base sequence of the messenger-RNA formed on it (Weiss & Nakamoto, 1961). Again, there is no opportunity for electrical impulses to interfere with the process to give a messenger-RNA of base sequence undetermined by its DNA template. The steric relationships necessary for synthesis on the template are absolutely specific for particular bases (Abrams, 1961).

After transfer to the ribosomes, the RNA serves as template for protein biosynthesis. The sequence of amino acids in any particular type of cell protein (including enzyme proteins) is determined directly by the base sequence of the RNA on which it forms, and, of course, indirectly by the base sequence of the DNA on which the RNA was formed (Speyer, Lengyel, Basilio, & Ochoa, 1962; Sueoka, 1961).

It is quite clear that the only way in which electrical impulses could disrupt any of these template mechanisms to give altered monomer sequences is by the formation of a mutagen. The absence of unusual bases in brain nucleic acids precludes this possibility. Moreover, even if a change in polymer base sequence was caused in a neurone by some outside influence, the effects would probably not have the results desired by Gaito (1961). Thus, unless the alteration in monomer sequence were a large one, which would certainly be disastrous for the cell due to the production of foreign protein, the only effect would be some small change in the amino acid sequence of a protein. To be of any consequence this would have to be an enzyme protein. If the change occurred at the active center, the enzyme would lose its catalytic properties and be useless to the cell. If the change occurred elsewhere in the protein molecule the chances are greatly against any significant alteration in properties (Fincham, 1960).

In summary, one may say that the hypothesis of brain polymer sequence alteration as the molecular basis of memory is unlikely because the known template mechanisms of polymer biosynthesis preclude any such changes, and even if such changes did occur they would, if large, lead to a complete disruption of cell activity, or, if small, to no significant change.

Experimental evidence in support of the polymer rearrangement hypothesis of memory has been published by several workers. This evidence is indirect and is concerned largely with the supposed role of RNAs as molecular memory-banks. It has been demonstrated that increases in total brain RNA occur with age (Cameron & Solyom, 1961), that the administration of structural analogs of purine bases affects rat maze learning (Dingman & Sporn, 1961), that treatment with ribonuclease affects the retention of conditional responses in regenerated planarians (Corning & John, 1961), and that activity produces alterations in the base ratios of RNA of brain cells (Egyhazi & Hyden, 1961; Hyden, 1959, 1961).

Considered as a whole, this evidence definitely demonstrates a role for RNA in the molecular processes basic to memory and learning. But, on logical grounds, it is necessary first to demonstrate that this role is not connected with the well-established function in protein biosynthesis, before postulating a second function involving nucleotide rearrangements that are improbable on chemical grounds. It is the purpose of the following discussion to propose that the molecular basis of memory involves changes

in enzyme concentration in brain cells, and that these changes necessarily involve the observed alterations in RNA metabolism.

It is a well-known principle of genetics that all somatic[1] cells contain the same genes. Each gene (consisting of DNA) controls the synthesis of a particular enzyme (Leslie, 1961). Hence every body-cell has the potential to make the same set of enzymes. But it is quite clear that there are large enzyme differences between cells. Some lack enzymes possessed by others, while the concentration of a particular enzyme varies widely between cells of different tissues (Dixon & Webb, 1958). It is apparent that the genes of some cells are inactive; in the sense that they are not participating in protein biosynthesis to any significant extent; while others are not functioning at their maximum rate. There must exist within cells a mechanism (or mechanisms) to initiate and control the rate of DNA, messenger-RNA, and messenger-RNA-protein syntheses. The chemical basis of these mechanisms is still obscure, but many enzymes are known to be induced by the presence of their substrates (Dixon & Webb, 1958). It is possible that the presence of substrate within a cell initiates synthesis, while the concentration of substrate determines the rate at which a particular enzyme is produced. This mechanism could be basic to the biochemical basis of memory and could operate in the following manner.

The synapses of the brain are polyfunctional, in that several axons terminate at the dendrites on any particular neurone. For this neurone to transmit an impulse it must be stimulated. The stimulus, which is chemically transmitted, can come either from a single axon, or from a combination of axons discharging together. Whether or not the impulse is transmitted depends upon whether the concentration of released transmitter substance reaches the critical level to stimulate the next neurone. Now, the amount of transmitter substance manufactured and stored by a neurone must depend upon several factors: including the activity within the cell of the biosynthetic enzyme system for the substance. If the enzymes of this system are inducible, the exchange of metabolites and demands for transmitter substance on repeated stimulation will increase the neurone's ability to manufacture the substance. Hence on the first few stimulations a neurone may not release sufficient transmitter substance to produce a postsynaptic response, but the amount of substance released on subsequent occasions will be sufficient.

If this hypothesis is incorporated with the concept of polyfunctional synapses and interconnecting neurone pathways in the brain, it becomes clear that an increased amount of transmitter substance released by any axon following repeated stimulation could be the major factor in establishing and maintaining the particular nervous mechanisms of memory and learned behavior.

Experimental investigation of the hypothesis could be attempted by a determination of whether the enzyme systems for the biosynthesis of nerve transmitter substances are inducible, and whether there is any increase in the amount of transmitter substance released by an axon following repeated stimulation. The recent reports of variations in cholinesterase activity with age, strain, individuals, and after learning (Aprison & Himwich, 1954; Bennett, Krech, Rosenzweig, Karlsson, Dye, & Ohlander, 1958; Bennett, Rosenzweig, Krech, Karlsson, Dye, & Ohlander, 1958; Krech, Rosenzweig, & Bennett, 1956, 1959; Krech, Rosenzweig, Bennett, & Longueil, 1959) in the brain of rats, could well be changes induced by increased amounts of the transmitter substance acetylcholine, which is being formed by neurones in larger quantities due to increased choline acetylase activity as suggested by the above hypothesis.

It will be seen that the hypothesis suggested here, that the molecular basis of memory and learning is basically induced changes in the enzyme systems involved in the formation and degradation of nerve transmitter substances, unlike the hypothesis of nucleotide sequence rearrangement, makes no proposal contrary to

current biochemical views. Moreover, increased concentration of specific enzymes necessarily implies increased activity of the RNA-dependent biosynthetic systems for these enzymes. In fact, certain definite changes in RNA metabolism are predicted by the hypothesis. These are outlined below.

Increases in enzyme concentrations imply increases in RNA. Considering the central nervous system as a whole, learning implies the continual establishment of new interconnecting neurone pathways. Hence a continual increase in brain RNA with age is to be expected. The increase in any particular neurone will occur relatively rapidly, but the increase within any tissue mass will be slow.

It is clear that the administration of any compound to an animal that will interfere with the RNA-dependent biosyntheses in neurones will affect learning and memory. Azaguanine is such a compound. It is a structural analog of guanine (a purine base that is a major constituent of both DNA and RNA). It is to be predicted that the presence of this analog in brain cells will disrupt the inductive increase in enzymes, due to an inhibition of the DNA- and RNA-dependent processes necessary for enzyme biosynthesis.

The hypothesis predicts increases in those molecular species of RNA required for the biosynthesis of enzymes involved in the metabolism of nerve transmitter substances. Other types of RNA will not be altered. Now every molecular species of RNA probably has a different ratio of nitrogenous bases. Consequently, prior to induction the total RNA of a cell will have a particular base ratio that is the mean of all the individual RNA base ratios. After induction some types of RNA have increased. Hence the base ratio of total RNA will now be different.

It will be seen that the changes in RNA metabolism predicted by the enzyme induction hypothesis are just those that have been experimentally reported and claimed to support the nucleotide sequence hypothesis. It is clear that this evidence cannot be used to decide between these two alternate hypotheses.

One further piece of experimental evidence will be considered. It has been reported by Corning and John (1961) that planarians conditioned to avoid light can be cut in two and that after regeneration both halves retain the conditional response. However, if the planarians are allowed to regenerate in water containing small amounts of ribonuclease (an enzyme catalyzing the hydrolysis of RNA), only organisms regenerated from "heads" retain the response.

These experiments are difficult to interpret on several grounds. First, ribonuclease is a protein and may not diffuse into the planarians' cells, or may not diffuse equally into all. Second, there is no evidence as to the quantity and distribution of neurones involved in the response. Thus on the enzyme induction hypothesis, assuming that the response involves some interaction between the head ganglia and the somatic neurones, certain levels of RNA are necessary to maintain induced levels of transmitter substances in particular neurone pathways. On treating the "tail" with ribonuclease these essential RNA levels are reduced to the preinduction concentrations. Hence the organism regenerated from this portion lacks the conditioned response. On either hypothesis, it is necessary for the ribonuclease not to destroy the neurone RNA in the head section. Consequently, this experiment also fails to provide an experimental test to distinguish between the two hypotheses.

In summary it may be said that the nucleotide rearrangement hypothesis of memory postulates undemonstrated properties of nucleic acids that seem incompatible with current knowledge, while the hypothesis fails to give any satisfactory account of how RNA-stored information could influence the nervous mechanisms of memory. The experimental evidence produced in favor of this hypothesis is equally compatible with the hypothesis that memory is basically dependent upon neurone pathways maintained by high levels of transmitter substances due to

induced biosynthetic enzymes. This enzyme induction hypothesis is also compatible with reported age and learning increases in brain cholinesterase levels: a phenomenon not predicted by the nucleotide rearrangement hypothesis.

REFERENCES

ABRAMS, R. Nucleic acid metabolism and biosynthesis. *Annu. Rev. Biochem.*, 1961, **30**, 165–188.

APRISON, M. H., & HIMWICH, H. E. Relation between age and cholinesterase activity in several rabbit brain areas. *Amer. J. Physiol.*, 1954, **179**, 502–506.

BENNETT, E. L., KRECH, D., ROSENZWEIG, M. R., KARLSSON, H., DYE, N., & OHLANDER, A. Cholinesterase and lactic dehydrogenase activity in the rat brain. *J. Neurochem.*, 1958, **3**, 153–160.

BENNETT, E. L., ROSENZWEIG, M. R., KRECH, D. R., KARLSSON, H., DYE, N., & OHLANDER, A. Individual, strain and age differences in cholinesterase activity of the rat brain. *J. Neurochem.*, 1958, **3**, 114–152.

CAMERON, D. E., & SOLYOM, L. Effects of ribonucleic acid on memory. *Geriatrics*, 1961, **16**, 74–81.

CORNING, W. C., & JOHN, E. R. Effect of ribonuclease on retention of conditioned response in regenerated planarians. *Science*, 1961, **134**, 1363–1365.

DINGMAN, W., & SPORN, M. B. The incorporation of 8-azaguanine into rat brain RNA and its effect on maze learning by the rat. An inquiry into the biochemical basis of memory. *J. psychiat. Res*, 1961, **1**, 1–11.

DIXON, M., & WEBB, E. C. *Enzymes*. London: Longmans, 1958.

EGYHAZI, E., & HYDEN, H. Experimentally induced changes in the base composition of the ribonucleic acids of isolated nerve cells and their oligodendroglial cells. *J. biophys. biochem. Cytol.*, 1961, **10**, 403–410.

FINCHAM, J. R. S. Genetically controlled differences in enzyme activity. *Advanc. Enzymol.*, 1960, **22**, 1–43.

FREESE, A. B. Transitions and transversions induced by depurinating agents. *Proc. Nat. Acad. Sci., Wash.*, 1961, **47**, 540–545.

GAITO, J. A biochemical approach to learning and memory. *Psychol. Rev.*, 1961, **68**, 288.

HYDEN, H. Biochemical changes in glial cells and nerve cells at varying activity. In, *Proceedings of the international congress of biochemistry: Biochemistry of the central nervous system*. Vol. III. London: Pergamon, 1959, Pp. 64–89.

HYDEN, H. Satellite cells in the nervous system. *Scient. Amer.*, 1961, **205**(6), 62–70.

KORNBERG, A. Biologic synthesis of deoxyribonucleic acids. *Science*, 1960, **131**, 1503–1508.

KRECH, D., ROSENZWEIG, M. R., & BENNETT, E. L. Dimensions of discrimination and the level of cholinesterase activity in the cerebral cortex of the rat. *J. comp. physiol. Psychol.*, 1956, **49**, 261–268.

KRECH, D., ROSENZWEIG, M. R., & BENNETT, E. L. Correlation between brain cholinesterase and brain weight within two strains of rats. *Amer. J. Physiol.*, 1959, **196**, 31–32.

KRECH, D., ROSENZWEIG, M. R., BENNETT, E. L., & LONGUEIL, C. L. Changes in brain chemistry of the rat following experience. Paper read at American Psychological Association, Cincinnati, September 1959.

LESLIE, I. Biochemistry of heredity: A general hypothesis. *Nature*, 1961, **189**, 260–268.

SPEYER, J. F., LENGYEL, P., BASILIO, C., & OCHOA, S. Synthetic polynucleotides and the amino acid code. *Proc. Nat. Acad. Sci., Washington*, 1962, **48**, 63–68.

SUEOKA, N. Correlation between base composition of DNA and amino acid composition of protein. *Proc. Nat. Acad. Sci., Wash.*, 1961, **47**, 1141–1149.

WEISS, S. A., & NAKAMOTO, T. On the participation of DNA in RNA biosynthesis. *Proc. Nat. Acad. Sci., Washington*, 1961, **47**, 694–697.

(Received December 11, 1961)

Psychological Review
1962, Vol. 69, No. 6, 542–547

THE UNTENABILITY OF LUCE'S PRINCIPLE

WILLIAM W. ROZEBOOM

Wesleyan University

In the course of his instructive researches in the mathematics of psychophysical scaling, R. Duncan Luce (1959a; 1959b, p. 29) has recently proposed a methodological principle which, if true, would have revolutionary import not merely for psychophysics, but for empirical and theoretical research in all branches of science. Moreover, while Luce's Principle has met with something less than enthusiastic acceptance in all quarters (cf. Luce, 1959b, p. 90; Mosteller, 1958, p. 287), others have begun to draw upon it as a basis for their own work (cf. Galanter & Messick, 1961). Such applications are decidedly premature, however, for the Principle's methodological status is dubious at best. As is inevitable for a provocatively novel hypothesis, certain ambiguities and unresolved difficulties reside in Luce's initial account, and when these are made explicit, it becomes highly problematic whether an interpretation can be found for the Principle which escapes being either vacuously true or empirically false.

Operating within the context of Stevens' theory of scale-types (e.g., Stevens, 1951),[1] Luce has proposed that the class of transformations defined by a scale-type should impose limitations on the mathematical form of the functional relations in which a variable of that type can participate. Specifically, it is postulated (Luce, 1959b) that:

> A substantive theory [i.e., an assumed law] relating two or more variables and the measurement theories [i.e., assumptions about the scale-types which are appropriate] for these variables should be such that:
>
> 1. (*Consistency of substantive and measurement theories*) Admissible transformations [i.e., transformations belonging to the class defined by the scale-type] of one or more of the independent variables shall lead, via the substantive theory, only to admissible transformations of the dependent variables.
>
> 2. (*Invariance of the substantive theory*) Except for the numerical values of parameters that reflect the effect on the dependent variables of admissible transformations of the independent variables, the mathematical structure of the substantive theory shall be independent of admissible transformations of the independent variables (p. 85).

Before putting his Principle to work, Luce comments that "one can hardly question the consistency part of the principle," but adds that "the invariance part is more subtle and controversial." I shall propose, to the contrary, that if this verbal statement of the Principle is taken at face value, it is the "consistency" part which is strong and correspondingly untenable, whereas the "invariance" part is a weakened version of the former which need be no more than a mathematical truism.

What can be meant when, in the consistency part of his Principle, Luce speaks of a transformation of an independent variable as *leading to*, via the substantive theory, a transformation of the dependent variable? The wording suggests that the transformation of the one somehow compels or induces a transformation of the other, an interpretation which is strengthened when, in subsequent illustration, Luce refers to a transformation being "effected on" the dependent variable. This cannot be a mathematical compulsion, however, for there is nothing about a lawful relationship which requires us to rescale one variable when transforming another. Thus given the law, "The area of a circle in square inches is π times its squared radius in inches," if I elect to scale length in feet rather than inches, my law simply becomes "The area of a circle in square inches is 144π times its squared radius

[1] I should also like to voice my reservations about a number of the conclusions of this theory. However, the substance of these doubts must await another occasion.

in feet." (If in this example the rescaling of length is felt—as it should not be—to require a corresponding rescaling of area, we could instead take, e.g., the relation between a person's height and his age.) The apparent meaning of the "consistency" stricture comes clearer upon examination of Luce's proffered illustration:

> Suppose it is claimed that two ratio scales are related by a logarithmic law. An admissible transformation of the independent variable x is multiplication by a positive constant k, i.e., a change of unit. However, the fact that $\log kx = \log k + \log x$ means that an inadmissible transformation, namely, a change of zero, is effected on the dependent variable. *Hence the logarithm fails to meet the consistency requirement* [italics added] (1959b, p. 85).

So worded, the example does not make explicit what happens to the dependent variable; however, let y be a function of x by the relation $y = \log x$, and let x' ($=_{\text{def}} kx$) be the transformed scale of the independent variable. Then the cited mathematical fact about $\log kx$ implies that $\log x' = \log k + \log x$, or $y + \log k = \log x'$; so if we wish to find a rescaling, y', of the dependent variable which stands in the same relation to x' as that in which y stands to x, it has to be $y' = y + \log k$, which is inadmissible (says Stevens) for a ratio scale. As demonstrated by this example, therefore, what the consistency injunction entails is that if an admissible transformation is made of the independent variable, then there must be an admissible rescaling of the dependent variable which leaves the mathematical expression of the law invariant—i.e., if $y = \phi x$, and x' is an admissible rescaling of x, then there is an admissible rescaling, y', of y, such that $y' = \phi x'$. Other than this, it is difficult to see what sense could be made of the notion that transformation of the one variable "effects" a transformation of the other.

Before defending this interpretation against another which also needs to be considered, it is useful to state explicitly just what does happen to a lawful relationship, $y = \phi x$, under transformation of its variables.[2] Let T and U be identity preserving (i.e., one-one) transformations, admissible or otherwise, applicable to x and y, respectively. Then if x is transformed into x' by T and y is transformed into y' by U, we have $x' = Tx$ and $y' = Uy$, or $x = T^{-1}x'$ and $y = U^{-1}y'$. Hence the substantive theory

$$y = \phi x$$

rewritten in terms of x' and y', becomes

$$U^{-1}y' = T^{-1}x'$$

or

$$y' = U\phi T^{-1}x' \qquad [1]$$

Moreover, if it is *also* the case that $y' = \phi x'$, it follows from Equation 1 that $\phi x' = U\phi T^{-1}x'$, or $\phi Tx = U\phi x$. Let T_a be an *admissible* transformation of x, and U_a, similarly, an admissible transformation of y. Then the consistency claim of Luce's Principle apparently requires that if $y = \phi x$, the function ϕ must be such that for any admissible transformation T_a, there corresponds an admissible transformation U_a such that for all values of x,

$$\phi T_a x = U_a \phi x \qquad [2]$$

(It is easily seen that if the inverse of ϕ exists, this condition on ϕ is satisfied if and only if $\phi T_a \phi^{-1}$ is an admissible transformation of y.)

It has been proposed to me that what Luce intended the consistency part of his Principle to assert is only that if an admissible transformation, $x' = T_a x$, is made of the independent variable in $y = \phi x$, then there must exist some relationship ϕ' and an admissible transformation, $y' = U_a y$, of the dependent variable such that $y' = \phi' x'$. Then to derive Equation 2, the invariance part of the Principle must be construed to stipulate that U_a can always be chosen to set $\phi' = \phi$. But if this is all that the consistency claim were to imply, then it would have no restrictive force what-

[2] Since x may be construed as a vector, the analysis which follows applies to multivariate relationships as well as to laws involving only one independent variable.

soever—we can always take $\phi' = \phi T^{-1}$, in which case y requires no transformation at all—and hence a case could never arise, contrary to Luce's example cited earlier, in which the consistency requirement is violated. Moreover, as the invariance part of the Principle is worded, it is not strong enough to authorize Equation 2 because it insists not on *strict* invariance of ϕ under admissible transformations of the variables, but only on an invariance of some *structure* of ϕ. But the mathematical form (i.e., structure) of a function is hopelessly nonunique, and unless additional, as yet unspecified, conditions are imposed on this to-be-invariant form, we may take it to be a form common to any function of the class $\{U_o \phi T_o^{-1}\}$—which then leaves the invariance condition without any bite.

In any event, that Condition 2 is, in fact, the mathematical nugget which Luce (1959b) extracts from the ore of his Principle is demonstrated by his Table 1. Each entry in the "Functional Equation" column may be seen to be an instance of Equation 2. In view of the stiflingly strong consequences (e.g., Luce, 1959a, Table 2) of Luce's Principle, what can be said to justify it? Despite the undeniable mathematical charm of Equation 2, it seems to me that the answer must be an uncompromising "nothing whatsoever." It will be noticed that Luce himself does not actually present any argument to support his Principle, but in effect merely invites the assent of the reader's intuition. But intuition may well be gulled by the ambiguities in the verbal statement of the Principle into acceding to an innocuous reading of it which then imperceptibly slides into a not-so-innocuous version. This is especially likely for the consistency part of the Principle, which Luce insists can hardly be questioned even while at the same time holding it to imply at least *some* restriction on the form of a law. (But it has already been argued that the consistency claim is either vacuous or carries the full burden of Equation 2 by itself.) In any case, in my own methodological research I have seen too many initially self-evident theses wither away under careful probing to have much confidence in methodological intuition.

Pending cognitive support for Luce's Principle, then, how well does it stand the test of applicability to established laws? Well, as Luce himself points out, it is not difficult to find laws (see example below) which do *not* conform to it. Luce's comments on this faux pas carry the discussion into a rather different problem of vast proportions, but before pursuing his lead, we may pause to raise an eyebrow at the methodological status of a "principle" which is not binding. To be sure, the verbal statement of Luce's Principle does not say that all substantive laws *do* satisfy the Principle, but only that they *should* do so. It is most unlikely, however, that what was intended by this wording is merely that conformity to the Principle is *desirable*. Science is concerned with what is, not with what ought to be, and if the Principle is to have any scientific significance, it must be that Condition 2 says something about the forms which laws are *in fact* able to assume, and in light of which we may limit the forms to be considered when speculating about laws hitherto undiscovered. But if there are perfectly good laws which violate Equation 2, then it is difficult to see how the Principle could ever by itself impugn the legitimacy of any assumption about the form of a problematic law.

If Luce's Principle is to have any methodological force at all, therefore, there must exist additional criteria by which we can discriminate between laws which fall under the Principle's sovereignty and those which do not. Luce (1959b) has tentatively identified such a criterion: "All physical examples which have been suggested to me as counterexamples . . . have a common form: the independent variable is a ratio scale, but it enters the equation in a dimensionless fashion" (pp. 90 ff.). That is, the law can be put into a form in which the independent variable, x, is multiplied times a parameter, c, whose numerical value is determined by the choice of scale for x, so that any change in unit for x may be absorbed into c without

disturbing the rest of the law—i.e., the law can be written in form $y = \phi_e(x) = \psi(cx)$, which under transformation $x' = kx$, goes over into $y = \phi_{e'}(x')$, where $c' = c/k$. Luce suggests that we can avoid violation of Condition 2 in such cases by taking the "dimensionless" variable v ($=_{\text{def}} cx$) as our independent variable; then the law becomes $y = \psi(v)$, in which v has no admissible transformations and there are hence no restrictions imposed on ψ by Condition 2. Thus a more subtle version of Luce's Principle emerges: What is now postulated is not that Condition 2 is inviolable, but that the laws of nature are always such as to be *capable* of an expression which satisfies Equation 2.

Clearly crucial to this revised formulation of Luce's Principle, however, is the concept of a dimensionless variable. For if the method for satisfying Equation 2 in prima facie embarrassing cases is to search out a dimensionless expression of the independent variables, the restrictive force (if any) of the Principle cannot be comprehended until we understand what is involved in such a construction. So pregnant a topic as this amply merits discussion and controversy, which I will do my bit to provoke by injecting some irritants which, hopefully, may stimulate Luce and perhaps others to further production of methodological pearls.

It is my considered opinion that to render a variable dimensionless in the fashion brought to our attention by Luce amounts to nothing but a more or less arbitrary selection of one of the admissible scalings of that variable, and then working up an "absolute" interpretation for that scale. I contend that such an interpretation can be found for at least one admissible transformation of any variable, no matter what its scale type (in fact, I would argue that *every* scale can have an absolute interpretation forced upon it), and hence that there is no law on whose mathematical form the revised version of Luce's Principle could impose restrictions. To defend this thesis in full generality is impractical here, so I shall confine my remarks to ratio scales.

An example cited by Luce of a law which flouts Condition 2, but which can be brought to heel by dedimensionalizing it, is the law of radioactive decay. If q is the quantity, say in grams, of a certain radioactive substance S at any time t, say the number of seconds after the beginning of the Twentieth century, then

$$q = ae^{-bt} \qquad [3]$$

where a and b are empirical constants. So formulated, the dependent variable (q) is a ratio scale while the independent variable (t) is an interval scale, which apparently contravenes not only Condition 2 (cf. Luce, 1959b, Theorem 4), but also Luce's suggestion that the only violations of Equation 2 are ones in which the independent variables are ratio scales. However, it is debatable whether generalizations such as Equation 3, in which a time coordinate assumes the status of an independent variable, should be regarded as genuine laws (cf. Rozeboom, 1961, pp. 356f.). If t_0 is an arbitrary point in time on the t scale, and q_0 is the quantity (on the q scale) of S at time t_0, then Equation 3 may be rewritten as

$$q = q_0 e^{-b(t-t_0)}$$

or letting $d =_{\text{def}} t - t_0$,

$$q = q_0 e^{-bd} \qquad [4]$$

Equation 4, in which all the variables are ratio scales, relates the quantity (in grams) of substance S at the end of a given time interval jointly to the interval's duration (in seconds) and S's quantity (in grams) at the interval's beginning. We can also express Equation 4 as a law with only one independent variable by dividing both sides by q_0 and setting $p =_{\text{def}} q/q_0$. We then have

$$p = e^{-bd}, \qquad [5]$$

which shows what proportion of S at a given moment still remains d seconds later, and in which the dependent variable, p, is an absolute, or dimensionless, scale.

Neither Equation 4 nor Equation 5

yet satisfy Condition 2,[3] but if we now ask how long a "half-life," L_h, substance S has—i.e., L_h is the value of d for which $p = .5$ in Equation 5—we find that $L_h = (\ln 2)/b$, and hence that

$$q = (\tfrac{1}{2})^{d/L_h} \qquad [6]$$

Then letting d_h be the ratio of d to L_h—that is, if the time duration variable is rescaled with the half-life of S as the unit of measurement—we finally obtain

$$q = (\tfrac{1}{2})^{d_h} \qquad [7]$$

in which the independent variable is an absolute scale with no admissible transformations, and which is hence immune to any threat from Condition 2. This, then, is the sort of procedure which, suggests Luce, may be used to placate Condition 2 when an empirical law involving ratio scales (e.g., Equation 4) appears to disregard it.

It should be appreciated, however, that to describe a variable such as d_h in Equation 7 as dimensionless is somewhat misleading. The scale is just as much dependent upon a unit of measurement as it ever was; only now one particular unit has been singled out as a *preferred* unit, which determines an absolute scale of the variable in the sense (and merely in this sense) that there is one and only one scale which expresses each value of the variable as a multiple of this preferred unit. This can obviously be done whenever a preferred unit of measurement has been selected, and hence *any* ratio scale law, no matter what its mathematical structure, can in this fashion be put into a form which is compatible with Equation 2.

There is still one more gambit by which some restrictive force might conceivably be salvaged for Luce's Principle: If a distinction could be drawn between *natural* units of measurement on one hand, and merely conventional ones on the other, the definition of dimensionless scales could be limited to variables for which a natural unit is available. While it is my belief that "naturalness" in this sense is purely a psychological effect with little or no methodological significance, I will not argue the point here. Instead, I merely invite anyone who feels otherwise to state what unit is the "natural" one for dedimensionalizing the time-duration variable in the law of radioactive decay. There is nothing special about the half-life interval. For any proportion r, we could just as well take substance S's r-life, L_r, for our standard (i.e., L_r is the time required for S to decay to the proportion r of its initial amount), and write our law as

$$p = r^{d/L_r} \qquad [6']$$

or

$$p = r^{d_r}. \qquad [7']$$

where d_r is the L_r-scale of duration. Preference for $r = .5$ is wholly a matter of psychological convenience.

[3] Equation 4 violates Luce's Theorem 1 (1959b), while Equation 5 is especially interesting in that since its inverse exists and its dependent variable has no admissible transformations (other than the Identity transformation), Condition 2 requires that the same be true of its independent variable.

SUMMARY

The *strong* version of Luce's Principle, namely, that all laws must conform to Condition 2, is not only unsupported by reason, but is refuted by counter-examples. On the other hand, a modified version of the Principle which demands of a law only that it be equivalent to one which satisfies Equation 2 (or, somewhat less generally, that a law is exempt from Equation 2 only if its independent variables can be dedimensionalized) is vacuous insomuch as any law, no matter how seriously at odds with Equation 2 in its original form, can be coaxed into compliance with the weakened requirement. Until further clarifications and supports for Luce's Principle are forthcoming, then, it would be questionable strategy to limit speculations about possible laws to mathematical forms approved by Equation 2.

REFERENCES

GALANTER, E., & MESSICK, S. The relation between category and magnitude scales of loudness. *Psychol. Rev.*, 1961, **68**, 363–372.

LUCE, R. D. *Individual choice behavior: A theoretical analysis*. New York: Wiley, 1959. (a)

LUCE, R. D. On the possible psychophysical laws. *Psychol. Rev.*, 1959, **66**, 81–95. (b)

MOSTELLER, F. The mystery of the missing corpus. *Psychometrika*, 1958, **23**, 279–289.

ROZEBOOM, W. W. Ontological induction and the logical typology of scientific variables. *Phil. Sci.*, 1961, **28**, 337–377.

STEVENS, S. S. Mathematics, measurement and psychophysics. In S. S. Stevens (Ed.), *Handbook of experimental psychology*. New York: Wiley, 1951. Pp. 1–49.

(Received February 14, 1962)

Psychological Review
1962, Vol. 69, No. 6, 548–551

COMMENTS ON ROZEBOOM'S CRITICISMS OF "ON THE POSSIBLE PSYCHOPHYSICAL LAWS" [1]

R. DUNCAN LUCE

University of Pennsylvania

The publication of Rozeboom's (1962) criticisms of my paper concerning limitations on scientific laws of certain types (Luce, 1959), together with what we both feel are misinterpretations and misleading uses of those results, prompt me to attempt to clarify some of the issues. At the outset, let me admit that much of the confusion is my fault, due in part to an inadequate understanding of the problem and in part to a not wholly satisfactory presentation of the ideas and results. These failures have been aggravated by some readers who seem to have ignored the second half of that paper and, thereby, have oversimplified its content.

Rozeboom takes me to task for an ambiguous statement of the two criteria that were suggested as a possible "principle of theory construction"—despite these ambiguities, which he has discussed in detail, we both agree about the mathematical "nugget" to be extracted from them, namely his Equation 2. We both emphasize that there are physical laws that either fail to satisfy this equation or satisfy it vacuously, and he thus concludes, correctly I now believe, that it should never have been termed a "principle." We also both point out that whether Equation 2 has any bite in limiting the form of a law has something to do with the existence of dimensional parameters in the law. We seem to differ mainly on the emphasis to place upon this distinction. Because I believe that it is extremely important, I should like to make a few remarks about it.

Among the various lawful statements that can be found in any science, Rozeboom and I are restricting our attention to what may be called function laws—the assertion of a functional relation among several variables, each of which can be represented as a simple numerical scale with certain invariance properties dictated by its measurement theory. As a specific example, from which we will generalize, consider the familiar law $E = mc^2$, where E denotes energy, m denotes mass, and c the velocity of light in a vacuum, all measured in some consistent set of units. It will be convenient to think of this law in the form $E - mc^2 = 0$. The first thing involved in this physical law is the assertion that there exists a particular function of three variables, call it ψ, such that all physically realizable values of E, m, and c satisfy

$$\psi(E, m, c) = 0$$

But there is a good deal more intended than just this simple mathematical equation. For example, not all of the arguments in the equation are of the same type: they are constrained in different ways. We usually speak of E as the dependent variable, m as the independent variable, and c as a dimensional parameter. By these words we mean such things as this: that certain of the symbols can assume arbitrary (positive) values and that the law then constrains others to have particular values, that other symbols represent constants that cannot be experimentally manipulated, and that certain measurement transformations of some of the symbols can be carried out provided that the transformations are admissible according to the measurement theories for these variables and that certain variables and parameters are transformed in prescribed ways. In other words, the statement of a law involves not only the usually written equation, but also certain specifications about which values can be arbitrarily selected, about which variables can be

[1] This work was supported in part by National Science Foundation Grant G-17637 to the University of Pennsylvania.

freely transformed, and about which values and transformations are prescribed when these choices are made.

To generalize from the example, we distinguish three kinds of symbols that enter into our functional relation (the distinctions among them are rendered more precise later): There are $m + 1$ variables, one of which is designated as

1. the dependent variable, y, and it assumes values in a set Y of real numbers, and the other $m \geqslant 1$ are designated as

2. independent variables, x_i, and they assume values in, respectively, sets X_i of real numbers; in addition, there are

3. $n \geqslant 0$ dimensional parameters, a_j.

Associated with each of these quantities is a group of admissible transformations. In the case of a variable, the group is determined by its measurement theory whereas, in the case of a parameter, it is simply included as part of the statement of the law. We denote these groups by $\mathbf{Y}$, $\mathbf{X}_i$, and $\mathbf{A}_i$. The more common transformation groups found in science include the affine group (multiplication by a positive constant), the full linear group, and the group of strictly monotonic increasing transformations—corresponding, respectively, to what are called ratio, interval, and ordinal scales.

Once the variables, parameters, and transformation groups are listed, the statement of a law then involves two distinct assertions. First, a particular function ψ of $1 + m + n$ variables is given which has the property that for any physically realizable choice of $x_i \epsilon X_i$, $i = 1, 2, \cdots, m$, any physically realizable value $y(x_1, \cdots, x_m) \epsilon Y$ of the dependent variable satisfies

$$\psi[y(x_1, \cdots, x_m), x_1, \cdots, x_m, a_1 \cdots, a_n] = 0$$

Second, for any set of (admissible) transformations $T_i \epsilon \mathbf{X}_i$, $i = 1, 2, \cdots, m$, of the independent variables, there exists an (admissible) transformation $U(T_1, \cdots, T_m) \epsilon \mathbf{Y}$ of the dependent variable and a set of (admissible) transformations $S_j(T_1, \cdots, T_m) \epsilon \mathbf{A}_j$, $j = 1, 2, \cdots, n$, of the dimensional parameters such that

$$\psi(y, x_1, \cdots, x_m, a_1, \cdots, a_n) = 0$$

implies

$$\psi[U(T_1, \cdots, T_m)y, T_1x_1, \cdots, T_mx_m, S_1(T_1, \cdots, T_m)a_1, \cdots, S_n(T_1, \cdots, T_m)a_n] = 0.$$

The first part of such a law, the function relating the variables, is always stated explicitly, but the second part is often implicit. It is simply a convention that scientists know and take for granted. For example, if a decay law for a particular radioactive material is

$$q = q_0 e^{-0.14t}$$

when the time t since a certain event is measured in seconds, then everyone knows that if we measure time in hours we must change the time constant from 0.14 to $(0.14)(360) = 50.4$.

Despite the fact that they are often left implicit, these transformation properties are important and must be made explicit in a full statement of a law. Indeed, one of the things that distinguishes theoretical science from a collection of simple numerical assertions are the relations among admissible transformations. In addition, the relations among transformations make the distinctions among the arguments of a law more than just linguistic conventions. Note that the independent variables are the arguments of ψ whose values in the first part and whose transformations in the second are selected arbitrarily; the value and transformation of the dependent variable are both determined via the law by those of the independent variable; and the value of a dimensional parameter is independent of the values of the variables and its transformation is determined by those applied to the independent variables. It is often possible to restate the law so that a different variable is treated as dependent; but this is not always possible, as for example, when the dependent variable is a periodic function of the independent ones.

Now consider a law (such as $F = ma$) that includes no dimensional parameters, and suppose that the solution y of

$$\psi(y, x_1, \cdots, x_m) = 0$$

is unique, i.e., there is a function ϕ of m variables such that

$$y = \phi(x_1, \cdots, x_m)$$

solves the above equation. Then the second part of the law can be rewritten as: for any $T_i \epsilon \mathbf{X}_i$, $i = 1, 2, \cdots, m$, then there exists $U(T_1, \cdots, T_m) \epsilon \mathbf{Y}$ such that if

$$y = \phi(x_1, \cdots, x_m)$$

then

$$U(T_1, \cdots, T_m)y = \phi(T_1x_1, \cdots, T_mx_m)$$

Substituting the first of these equations into the second, we obtain the following restriction on the function

$$U(T_1, \cdots, T_m)\phi(x_1, \cdots, x_m) = \phi(T_1x_1, \cdots, T_mx_m)$$

which is Rozeboom's Equation 2 for m independent variables; it includes the several special cases investigated in Luce (1959). For ratio and interval scales and $m = 1$, it was shown that, aside from numerical constants, there are very few possible solutions ϕ to this functional equation.

When dimensional parameters are admitted, the resulting equation is very much less restrictive. Specifically, with one independent variable and one dimensional parameter, the same argument leads to

$$U(T)\phi(x, a) = \phi[Tx, S(T)a]$$

which has many possible solutions. For example, suppose that both $y = \phi(x, a)$ and x are variables measured on ratio scales, and suppose that f is an arbitrary function of one variable. A solution to this equation is $\phi(x, a) = f(xa)$ provided that when x is transformed by $Tx = kx$, then U is the identity transformation and $S = T^{-1}$, i.e., $Uy = y$ and $Sa = a/k$.

Two conclusions seem justified. First, Rozeboom is correct in saying that the equation that results when there are no dimensional parameters is much too restrictive to be called a principle of theory construction. Second, and this I feel to be the main thrust of both my 1959 paper and this note, psychologists (as well as other scientists) either are restricted to a very few possible types of laws (in the case of one independent variable, to those given in my 1959 paper when the variables are measured on either ratio or interval scales) or they cannot avoid including dimensional parameters in the statement of their laws. Not only must these parameters be written explicitly, but a careful prescription should be included about how they transform as a function of admissible transformations of the independent variables, for otherwise the statement of the law is incomplete. Following the practice of physicists, the transformational aspects of laws are often not made explicit. This does not seem to create serious problems when the transformations are limited to the affine group (ratio scales), but matters are much less obvious when interval and weaker scales are involved. In these cases, which often seem to arise in psychology, it is essential that careful attention be given to the transformation properties of the variables and parameters.

Finally, the problem remains whether or not the number of dimensional parameters involved in the statement of a law is of any inherent importance. Some physicists seem to feel that such parameters are undesirable and that their total number in a science should be held to a minimum. My earlier results strongly suggest that the minimum cannot possibly be zero. Although psychologists have hardly begun to face these problems, inevitably they will. At that time, it would be useful to have a better understanding of the role of dimensional parameters than we have now. In physics they have meaning—the velocity of light, the density of a substance, the viscosity of a fluid, the gravitational constant, etc.—and they recur in various combinations in different laws. As yet, we have little of this in psychology.

SUMMARY

It is pointed out that certain scientific laws assert not only a functional relation among variables and dimensional parameters, but that they also state what transformations on the dependent variable and the parameters are necessary for the same function to be satisfied when admissible transformations are applied to the independent variables. It is shown when there are no dimensional parameters, these transformation conditions are very restrictive but that they are not when one or more dimensional parameters are present. The conclusion is that psychological laws will in general involve dimensional parameters and that it is necessary to state carefully the transformational aspects of the law.

REFERENCES

LUCE, R. D. On the possible psychophysical laws. *Psychol. Rev.*, 1959, **66**, 81–95.

ROZEBOOM, W. W. The untenability of Luce's principal. *Psychol. Rev.*, 1962, **69**, 542–547.

(Received May 9, 1962)

Psychological Review
1962, Vol. 69, No. 6, 552

COMMENT

WILLIAM W. ROZEBOOM

Wesleyan University

I am grateful to the Editor for allowing me this opportunity to express my pleasure at Luce's gracious acceptance of my criticisms of his earlier paper, and to record my full endorsement of his present conclusions. Briefly, the point is simply this: If a law is stated without stipulation of one particular scale for each of the variables involved, and if, moreover, it is intended that alternative scaling of these variables be countenanced, then it is also necessary to make explicit provision in the law's statement for the manner in which the contemplated transformations will affect it. In particular, if it is assumed that the function relating the variables is invariant under these transformations—and I think that Luce is quite right in his judgment that this assumption is normally made, though I am much less convinced of its methodological necessity—then it will usually be necessary to include scale parameters in the law whose meanings and transformational behavior also require study and clarification. Thanks to Luce, explicit awareness of this need has now been entered in the mainstream of psychological literature, and a few more bolts on psychology's shaky methodological scaffolding have begun to tighten.

(Received June 8, 1962)

ERRATA

In the article by Frank Restle entitled "The Selection of Strategies in Cue Learning" (*Psychol. Rev.*, 1962, **69**, 329–343), the first equation at the top of page 333 should read:

$$\cdot \frac{N_c + N_i - N_{n,1} - \cdots - N_{n,j-2} - N_{n,j-1}}{N_c + N_i - N_{n,1} - \cdots - N_{n,j-2} - N_{n,j-2}} \cdot \frac{N_{n,j}}{N_c + N_i - N_{n,2} - \cdots - N_{n,j-1} - N_{j-1}}$$

In the Errata on page 378 of the same issue, a line was omitted. It should read as follows (the omitted line appearing in italics):

"In the table of contents on the front cover of the May 1962 issue, *the title of the article by Joan E. Jones should read as follows:* '*Con*tiguity and Reinforcement in Relation to CS-UCS Intervals in Classical Aversive Conditioning'."

THE BRITISH JOURNAL OF PSYCHOLOGY

Edited by BORIS SEMEONOFF

Vol. 53 Part 3 August, 1962 20s. net

EDITORIAL. The B. J. P. in Jubilee Year.

T. H. PEAR. Personalities in the early days of the British Psychological Society.

CYRIL BURT. The concept of consciousness.

IAN M. L. HUNTER. An exceptional talent for calculative thinking.

L. B. BROWN. A study of religious belief.

H. BLAIR HOOD. An experimental study of Piaget's theory of the development of number in children.

A. E. M. SEABORNE. Group influence on the perception of ambiguous stimuli.

H. J. EYSENCK. Conditioning and personality.

EDWARD A. FEIGENBAUM and HERBERT A. SIMON. A theory of the serial position effect.

A. D. B. CLARKE and MARGARET COOKSON. Perceptual-motor transfer in imbeciles: a second series of experiments.

BOBBY J. FARROW and JOHN F. SANTOS. Changes in autokinetic perception as a function of the transfer of conditioning effects.

PUBLICATIONS RECENTLY RECEIVED.

The subscription price per volume, payable in advance, is 60s. net (post free).

Subscriptions may be sent to any bookseller or to the

CAMBRIDGE UNIVERSITY PRESS
Bentley House, Euston Road, London, N. W. 1

ZEITSCHRIFT FÜR EXPERIMENTELLE UND ANGEWANDTE PSYCHOLOGIE

Organ der Deutschen Gesellschaft für Psychologie

Herausgeber für Deutschland

PROF. DR. PH. LERSCH, München; PROF. DR. F. SANDER, Bonn

Herausgeber für Österreich	Herausgeber für die Schweiz	Redaktion
PROF. DR. H. ROHRACHER, Wien	PROF. DR. R. MEILI, Bern	DR. C. J. HOGREFE, Göttingen

Band IX/1

DIETER KLEBELSBERG, **Unterschiedliche Auswirkungen einer Blutalkoholkonzentration von 1, 0 Promille bei einzelnen psychischen Komponenten der Fahrtüchtigkeit**

ARNULF RÜSSEL, **Belastung und Ermüdung an unterschiedlichen Arbeitsplätzen**

KARL HEINZ SEIFERT, **Kompensatorische Leistungsmöglichkeiten Mindersinniger im Bereich der Haptik**

KONRAD STEININGER, **Aktuelle luftfahrtpsychologische Probleme**

LORÁND SZALAY, **Untersuchungen zur semantischen Struktur der Zeitwörter**

All articles are provided with summaries in English and French.
The journal appears quarterly and contains 160 pages per number

Subscription is DM 64,—per Volume

VERLAG FÜR PSYCHOLOGIE—DR. C. J. HOGREFE—GÖTTINGEN
(Brentanoweg 10, Germany)